UNIVERSITÀ
DEGLI STUDI
FIRENZE

PhD Program in Smart Computing
Dipartimento di Ingegneria dell'Informazione (DINFO)

Deep Domain Adaptation for Pedestrian Detection in Thermal Imagery

Kieu My

Dissertation presented in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Smart Computing

PhD Program in Smart Computing
University of Florence, University of Pisa, University of Siena

Deep Domain Adaptation for Pedestrian Detection in Thermal Imagery

Kieu My

Advisor:

Prof. Andrew D. Bagdanov

Head of the PhD Program:

Prof. Paolo Frasconi

Evaluation Committee:
Dr. Bogdan Raducanu, *Computer Vision Center, Barcelona, Spain*
Dr. Angel Sappa, *ESPOL CIDIS, Ecuador*

XXXIII ciclo — October 2021

To XXXXX

Acknowledgments

First of all, I would like to give the biggest thanks to my patient and supportive supervisor, Professor Andrew David Bagdanov, whom we called a familiar name: Andy. He gave me an excellent opportunity to work on the subject of Computer Vision under his kind supervision. He supported me wholeheartedly from the very first date when I submitted the scholarship proposal. Then a warm welcome at Firenze's railway station with detailed guidance on all aspects to start a new life when I just landed in Italy. And until this day, his mentoring remains vital and essential for both my personal and academic journey.

With many thanks to his comprehensive and dedicated instruction, from a computer vision newbie, I achieved the best student paper honorable mention award for my first publication, followed by other publications in the top-tier conferences in the world. Also, completing the Ph.D. in such an unusual time of Covid-19 would not be easy without Andy's understanding, sympathy, and flexibility, allowing me to learn, adapt, and grow professionally and personally. Overall, the three-year experience in my Ph.D. program is one of the most meaningful journeys in my life, and it probably would not have been possible without Andy. I sincerely appreciate all his help.

I would like to express my gratitude to Professor Marco Bertini who gave me many thoughtful advice and discussions and brought me into the fantastic project with RFI company. Marco is not only a great lab director, a member of my supervisory committee in three years, but also a great colleague in both research and industry area.

I would like to thank all of my colleagues at MICC, who are both talented in research and very friendly in social life. We have enjoyed many parties and dinners together. I am delighted to become a member of such a great lab of excellent researchers. Especially, I would like to give a big thanks to Dr. Claudio Baecchi who left a massive footprint on my work and my way of thinking. He is always an expert in my mind, and he is the main source of general computer science knowledge, and the first place my compass turns to when I mess things up, which is quite often. The next thought goes to my two "Ph.D. Comrades" Riccardo Del Chiaro and Matteo Bruni. We started this journey at the same time; we went to lectures and courses together and went through a lot of amazing experiences side by side. Specifically, I want to thank Riccardo for fully supporting me throughout weekly discussions with Andy and being especially helpful in technical issues.

I would also like to thank all the remaining great seniors in the lab: former lab director Professor Alberto Del Bimbo for his masterful skills and thoughtful advice, Professor Pietro Pala, and Professor Stefano Berretti for an interesting course on Manifold learning. Thanks to Dr. Lorenzo Seidenari for many engaging discussions

and useful advice on my presentation and two constructive courses on the Ph.D. program: Introduction to Machine Learning and GANs. Thanks to Dr. Tiberio Uricchi for his instruction and lesson in the first course of my Ph.D. program. A special thanks to Dr. Federico Becattini for all of his support for everything I need. And thanks to Dr. Roberto Caldelli, Dr. Federico Pernici, Dr. Leonardo Galteri, Dr. Claudio Ferrari, Dr. Paolo Mazzanti, Dr. Francesco Turchini, Dr. Andrea Ferracani for all their kind support and positive spirit. For junior Ph.D. colleagues, I would like to thank Lorenzo Berlincioni, Pietro Bongini, and Simone Ricci for their collaboration during my last year. We went through summer school, conference, collaborated on a paper and having lunch every day together. It has been a great pleasure working with them.

Stepping outside the lab, first, I would like to thank Dr. Joost van de Weijer (Barcelona, Spain) for being on my supervisory committee for three-years and providing helpful advice and comments. Thanks to Professor Paolo Frasconi for the Meta-Learning course, where his teaching style impressed me and his other big support as the department director. I also want to thank Simona Altamura for supporting me through Ph.D. administrative procedures.

I also want to thanks Dr. Bogdan Raducanu (Barcelona, Spain) and Dr. Angel Sappa (ESPOL CIDIS, Ecuador) for being my external reviewers of my Ph.D thesis and providing me useful feedback.

This 3-year Ph.D. journey would not have been possible without the great working environment and financial support from the University of Florence, to whom I am sincerely grateful.

Last but not least, one of the most important people besides me in the last decade is my wife, Le Hai Yen Tran. Although she works in a different field, as a consultant for the World Food Program, she understands and encourages my visions and dreams. We talked to each other every day, traveled throughout Europe, and had a great time together in Italy in the last three years. And there is no way to thank my family: my mother and two other younger sisters for everything they have done for me. Especially Ny Kieu, she always encourages me from my bachelor's degree until my Ph.D. program.

Every single word in this thesis is for every one of you who taught, inspired, and believed in me that I could have made it this far. Your love, patience, and encouragement are more valuable than you could ever imagine.

Thank you all for everything.

Abstract

Pedestrian detection is a core problem in computer vision due to its centrality to a range of applications such as robotics, video surveillance, and advanced driving assistance systems. Despite its broad application and interest, it remains a challenging problem in part due to the vast range of conditions under which it must be robust. In particular, pedestrian detectors must be robust and reliable at nighttime and in adverse weather conditions, which are some reasons why thermal and multispectral approaches have become popular in recent years. Moreover, thermal imagery offers more privacy-preserving affordances than visible-spectrum surveillance images. However, pedestrian detection in the thermal domain remains a non-trivial task with much room for improvement.

Thermal detection helps ameliorate some of the disadvantages of RGB detectors – such as illumination variation and the various complications of detection at nighttime. However, detection using only thermal imagery still faces numerous challenges, and overall lack of information in thermal images. Thermal images are typically low-resolution, which in turn leads to more challenging detection of small pedestrians. Finally, there is a general lack of thermal imagery for training state-of-the-art detectors for thermal detection. The best pedestrian detectors available today work in the visible spectrum.

In this thesis, we present three new types of domain adaptation approaches for pedestrian detection in thermal imagery and demonstrate how we can mitigate the above challenges such as privacy-preserving, illumination, lacking thermal data for training, and lacking feature information in thermal images and advance the state-of-the-art. Our first contribution is two *bottom-up domain adaptation* approaches. We first show that simple bottom-up domain adaptation strategies with a pre-trained *adapter* segment can better preserve features from source domains when doing transfer learning of pre-trained models to the thermal domain. In a similar vein, we then show that bottom-up and *layer-wise* adaptation consistently results in more effective domain transfer. Experimental results demonstrate efficiency, flexibility, as well as the potential of both bottom-up domain adaptation approaches.

Our second contribution, which addresses some limitations of domain adaptation to thermal imagery, is an approach based on task-conditioned networks that simultaneously solve two related tasks. A detection network is augmented with an auxiliary classification pipeline, which is tasked with classifying whether an input image was acquired during the day or at nighttime. The feature representation learned to solve this auxiliary classification task is then used to *condition* convolutional layers in the main detector network. The experimental results of task-conditioned domain adaptation indicate that task conditioning is an effective way to balance the trade-off between the effectiveness of thermal imagery at night and its weaknesses during the day.

Finally, our third contribution addresses the acute lack of training data for thermal domain pedestrian detection. We propose an approach using GANs to generate synthetic thermal imagery as a type of generative data augmentation. Our experimental results demonstrate that synthetically generated thermal imagery can be used to significantly reduce the need for massive amounts of annotated thermal pedestrian data.

Pedestrian detection in thermal imagery remains challenging. However, in this thesis, we have shown that our bottom-up and layer-wise domain adaptation methods – especially the proposed task-conditioned network – can lead to robust pedestrian detection results via using thermal-only representations at detection time. This shows the potential of our proposed methods not only for domain adaptation of pedestrian detectors but also for other tasks. Moreover, our results using generated synthetic thermal images also illustrate the potential of generative data augmentation for domain adaptation to thermal imagery.

Contents

Chapter 1

Introduction

Pedestrian detection is a core problem in computer vision due to its centrality to a range of applications such as robotics, video surveillance, and advanced driving assistance systems. Despite its broad application and interest, it remains a challenging problem in part due to the vast range of conditions under which it must be robust. In particular, pedestrian detectors must be robust and reliable at nighttime and in adverse weather conditions, which are some reasons why thermal and multispectral approaches have become popular in recent years. Moreover, thermal imagery offers more privacy-preserving affordances than visible-spectrum surveillance images. However, pedestrian detection in the thermal domain remains a non-trivial task with much room for improvement. This chapter gives an overview of the main challenges and approaches to pedestrian detection in the visible spectrum, in multispectral imagery, and using only thermal images.

1.1 Object and Pedestrian Detection

Object detection is one of the most important problems in Computer Vision and its applications are ubiquitous. In general, object detection is a combination of *object recognition* and *object localization*. The object recognition task is to estimate the likelihood of a semantic object instance in an image (e.g., the likelihood of there being a *dog* in an image or not). On the other hand, object localization refers to identifying and outputting the coordinates of known semantic objects in an input image. Object detection is the task of *detecting* and *localizing* semantic objects of known classes in images or video.

Object detection is one of the cornerstones of image understanding. It forms the basis for solving complex or high-level vision tasks such as object tracking, segmentation, image captioning, event detection, activity recognition, and scene understanding. Object detection supports various range of applications, including consumer electronics, autonomous driving, robot vision, security, human-computer

interaction, content-based image retrieval, intelligent video surveillance, and augmented reality. For a broad survey of deep learning methods for object detection, please see the review by Liu et al. (2019a).

Object detection methods typically fall into one of two categories: traditional machine learning-based approaches or deep learning-based approaches. Most of the early object detection algorithms were based on traditional machine learning-based techniques and hand-crafted features. There is a vast literature on classical approaches to object detection using hand-crafted features like SIFT or HOG and classifiers like Support Vector Machines (SVMs). According to Zou et al. (2019), the progress in object detection research was slow during 2010-2012 with a small number of proposed methods as the performance of hand-crafted features became saturated.

In 2012, however, we saw the resurgence in interest in deep learning, which pushed the object detection problem to renewed prominence. These new advances based on were made possible by the combination of a number of factors:

- **Deep Learning**: The renaissance of deep neural networks (DNNs), in particular Convolutional Neural Networks (CNNs) which can learn high-level, end-to-end representations of visual data.

- **Big data**: More and more data coming from the internet, mobile devices, and social network created an inexhaustible source of data for learning very large deep models.

- **Dedicated GPUs**: Graphics Processing Units (GPUs) significantly reduce the computational cost of deep learning, which was the final factor in the renewed interest in Deep Learning.

Many Convolutional Neural Network architectures have been used as the main backbone for state-of-the-art methods for object detection, including Region Proposals Networks (R-CNN) (Girshick et al., 2014), Faster R-CNN (Ren et al., 2015), the Single Shot MultiBox Detector (SSD) (Liu et al., 2016b), and You Only Look Once (YOLO) (Redmon and Farhadi, 2018) – to name but a very few.) These network architectures are used not only as object detectors but also as pre-trained models for transfer learning or domain adaptation to specific tasks. For example, YOLO version 2 by Redmon and Farhadi (2017) could detect around 9000 objects and is used as a pre-trained model for fine-tuning to many particular tasks such as car and pedestrian detection.

Pedestrian detection is another essential topic in computer vision because of its centrality in safety and security applications such as video surveillance, autonomous driving, robotics, criminal investigation, etc. Pedestrian detection has received significant attention from the research community over the past two decades with over

Figure 1.1: Example pedestrian detection results on RGB images from the KAIST dataset. There are a total of eight pedestrians on the left image, two pedestrians are occluded by others, and two pedestrians are missed due to insufficient illumination. There are a total of four pedestrians in the right image, with two of them being missed.

two thousand publications. According to the survey of Cao et al. (2020), with fewer than fifty publications on pedestrian detection every year before 2005, interest has gradually increased to around 300 publications every year for the last three years.

These numbers illustrate that pedestrian detection is still an active research problem in computer vision. Markit (2019) noted that there were an estimated 240 million installed video surveillance cameras worldwide in 2014. The continued need for detection and observation of humans in public spaces and the advent of autonomous driving systems promises to add many more cameras. For autonomous driving systems, detecting pedestrian must to be as accurate as possible. Figure 1.1 gives two examples of pedestrian detection results on RGB images from the KAIST Multispectral Pedestrian Detection Benchmark dataset (Hwang et al., 2015). The detection results are from a YOLOv3 detector fine-tuned on the KAIST training set. These images illustrate, first of all, that despite the sunny conditions there are still many challenging problems for pedestrian detection. There are a total of eight pedestrians in the left image, some of them are clearly visible and some are quite difficult to see – two are occluded by other pedestrians, and two pedestrians are missed due to illumination. Similarly, there are a total of four pedestrians in the right image, however two are not detected. Despite the significant improvement in pedestrian detection over the years, there remains a vast array of challenges that must be overcome to meet real-world application requirements.

1.2 Challenges of Pedestrian Detection in the Visible Spectrum

Despite the large number of detectors proposed in recent years, pedestrian detection in RGB images remains challenging due to the need to detect pedestrians accurately and quickly. In this section, we briefly discuss some of the challenges of pedestrian detection in RGB imagery.

1.2.1 Occlusion

Occlusion is one of the most challenging phenomena for pedestrian detection, even more so because it is also inevitable. There are many types of occlusion relevant to detection, but we can divide them into two broad categories: object occlusion and crowd occlusion. Object occlusion refers to pedestrians occluded by other objects such as vehicles, trees, or occluding accessories such as briefcases or umbrellas. Crowd occlusion, on the other hand, refers to pedestrians occluded by other pedestrians. Detecting occluded pedestrians is challenging because some parts of the body may be entirely missing when pedestrians are occluded. These missing visual features lead to severe degradation of pedestrian detectors if it does not handle properly. Occlusion was addressed by several works, for example Ouyang et al. (2016a) who learned a mutual visibility relationship between occluded pedestrians to mitigate problems arising from occlusions. Occlusion problems in detecting multiple persons was addressed by Hadi et al. (2015) with a vision-based model which used a fusion of depth and thermal images.

To sidestep the occlusion problem, especially when the size of objects is small, many state-of-the-art pedestrian datasets such as Caltech (Dollar et al., 2009) and the KAIST Multispectral Pedestrian Detection Benchmark (Hwang et al., 2015) do not consider large occluded cases as ground-truth for the detection task. Instead, they filter objects occluded by more than 50% from consideration entirely. Figure 1.2 illustrates the overlap and occlusion between pedestrians, which is the reason for many false-negative detection results. In this figure there are two pedestrians occluded by cars in the left and right images, and pedestrians are occluded by other pedestrians in a group in the middle image.

1.2.2 Varying illumination conditions

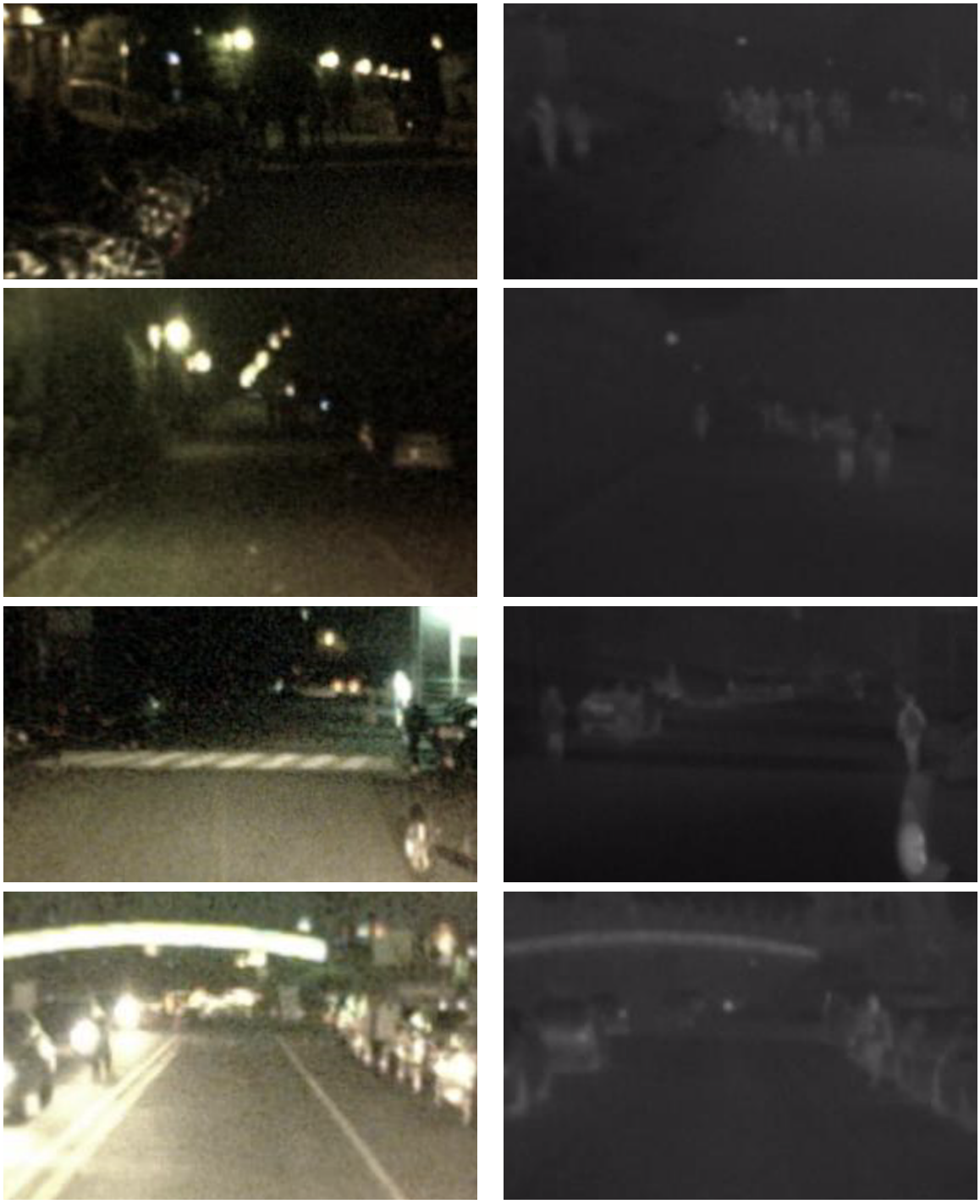
In the last decade, the majority of existing RGB detectors work acceptably on high-quality, reasonably controlled RGB images (Benenson et al., 2014; Zhang et al., 2016). However, many such detectors fail under illumination changes (e.g. nighttime) or adverse weather conditions such as rain or fog Li et al. (2018). According to

Figure 1.2: Examples of occluded pedestrians in the KAIST dataset. Two pedestrians are occluded by cars in the left and right images, while pedestrians are occluded by other pedestrians in the middle image.

Figure 1.3: Example of visible image (left) and thermal image (right) at nighttime. [Image from (Chen et al., 2020)].

Gawande et al. (2020), illumination variation is one of the most challenging problems for detecting pedestrians. Illumination might change due to lighting conditions, motion of the light source, reflection from bright surfaces, the effect of other light sources, or different times of the day. Figure 1.3 from Chen et al. (2020) shows an example of a visible spectrum image (left) and corresponding thermal image (right) at nighttime.

Figure 1.4 gives additional nighttime examples from the KAIST dataset. These illustrate the effect of illumination on pedestrian detection. In the left column, we see examples of insufficient lighting (the first three rows) and one example of too much light (the last row). It is difficult even for humans to discern all of the pedestrians in these images. In the corresponding thermal images in the right column of Figure 1.4, we see that pedestrians are significantly more identifiable.



(a) RGB Image (b) Thermal Image

Figure 1.4: Examples of pedestrian at nighttime with insufficient lighting (first three rows) and too much light (last row). Pedestrians are had even for humans to discern in the visible spectrum images, while they are significantly more evident in the corresponding thermal images.

1.2.3 Privacy preservation

We conclude our discussion of the challenges to pedestrian detection, not with another technological problem, but rather with a societal one. With the total number of installed video surveillance cameras already at 240 million worldwide in 2014 (Markit, 2019), and the advent of autonomous driving promising to add many more cameras – all detecting and observing humans in public spaces – citizens are naturally concerned that being observed violates their right to privacy. As demonstrated in Figure 1.3, thermal imagery offers advantages over RGB images, especially at nighttime. However, given that thermal sensors capture the radiant heat of objects, thermal imagery can provide clear object silhouettes and textures while offering *significant privacy-preserving affordances* not offered by visible spectrum imagery.

1.3 Pedestrian Detection in Thermal Imagery

Thermal imaging works as follows. All objects emit infrared energy (heat), which is known as a heat signature. The hotter an object is, the stronger radiation it emits. The sensor collects infrared radiation from objects in the scene and creates an image based on information about the temperature differences. A thermal camera is basically a heat sensor capable of detecting tiny temperature differences. Because objects are rarely exactly the same temperature as others around them, a thermal camera can image them distinctly in a thermal image. Thermal cameras were originally designed as a surveillance and night vision tool for the military. The work in Gade and Moeslund (2013) reviewed a vast array of applications based on thermal cameras such as detecting symptoms of animals without touching, checking agriculture and food quality, building inspection, gas detection, industrial applications, fire detection, medical analysis, as well as detection, tracking, and recognition of humans. According to Gade and Moeslund (2013), based on the wavelength spectrum of emission and absorption of infrared radiation between visible light and microwaves (0.7 - 1000 μm), the infrared spectrum can be divided into five spectral regions, including Near-infrared - NIR (0.7-1.4μm), Short-wavelength infrared - SWIR (1.4-3μm), Mid-wavelength infrared - MWIR (3-8μm), Long-wavelength infrared - LWIR (8-15μm), and Far-infrared - FIR (15-1000μm).

Thermal imagery, acquired by an infrared camera, might seem an ideal solution for the pedestrian detection task. It can solve some of the challenges of detection in RGB images (illumination variation, occlusion, privacy-preservation) by detecting infrared radiation of objects and generating an image based on that information Negied et al. (2015). Here we list a number of advantages pedestrian detection in thermal imagery has over detection in the visible spectrum.

1.3.1 Robustness to illumination variation

Because thermal cameras image objects by temperature and not by reflectance of visible light, thermal images can image pedestrians clearly under a range illumination conditions. Figure 1.3 from Chen et al. (2020) and figure 1.4 from the KAIST dataset (Hwang et al., 2015) illustrate examples of pairs of visible and thermal images at nighttime under insufficient lighting conditions. As we can see, thermal images can image pedestrians more clearly under insufficient illumination conditions. This is a crucial advantage for the pedestrian detection, and many works have used thermal images to compensate for inadequate illumination such as Guan et al. (2018), who used thermal and visible images in an illumination-aware multispectral DNN to learn multispectral human-related features under different illumination conditions. Li et al. (2019) also used thermal and visible images for their Illumination-aware Faster R-CNN method to perform multispectral pedestrian detection.

1.3.2 Robustness to occlusion

Even if pedestrians occluded by on another or by other thin surfaces, their temperature shape can appear clearly in thermal cameras. Many works have leveraged the advantages of thermal imagery for handling occlusion in pedestrian detection, such as Kristoffersen et al. (2016) who handled pedestrian occlusion problem using a stereo thermal camera, and Chen et al. (2019) who used Faster-RCNN and a region decomposition network to detect a wider range of pedestrian appearances including partial body poses and occlusions in thermal imagery, and Chen et al. (2020) who proposed an extension to Faster-RCNN for nighttime pedestrian detection in thermal images to deal with the occlusion problem under challenging illumination.

1.3.3 Privacy-preservation and thermal imagery

Detecting pedestrians using *only thermal images* is a potential solution for privacy preservation in video surveillance applications.* Figure 1.5 gives an example of cropped pairs of color and thermal images from the KAIST dataset (Hwang et al., 2015). As we can see from these examples, even in relatively low-resolution color images, persons can be readily identified. Meanwhile, thermal images still retain distinctive image features useful for detection while preserving privacy. Thermal imagery is privacy-preserving in the sense that person identification is difficult or impossible. Our approaches are partially motivated by the fact that thermal images guarantee some balance between security and privacy concerns.

*Note that all of the detectors we propose in this thesis are *thermal-only*, which we distinguish from *multispectral* detectors that use some combination of visible and thermal spectra and thus are not privacy-preserving.

Figure 1.5: Thermal imaging and privacy preservation. Shown are four cropped images from the KAIST dataset. On the left of each is the RGB image, to the right the crop from the corresponding thermal image. Note how persons are readily identifiable in visible spectrum images, but not in the corresponding thermal images. Although identity is concealed, there is still enough information in thermal imagery for detection.

1.4 The State-of-the-art in Pedestrian Detection

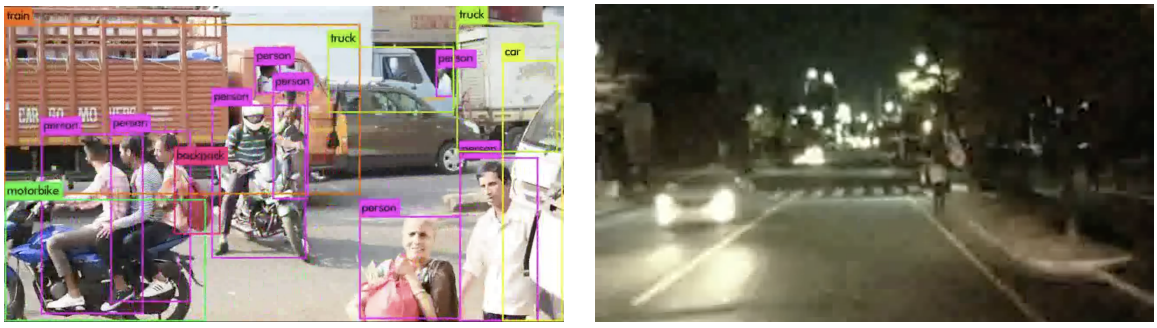
The state-of-the-art in pedestrian detection – in terms of both speed and accuracy – is in the RGB domain. The literature on pedestrian detection is vast and spans many decades. A complete review of he literature is beyond the scope of this thesis, but the interested reader should consult the excellent review by Zou et al. (2019) for a historical perspective, and the review by Liu et al. (2019a) for a treatment of recent deep learning-based approaches. Here we concentrate on the recent works most relevant to our contributions.

Today there are many state-of-the-art RGB detectors which are extremely fast and accurate – such as YOLO by Redmon and Farhadi (2018), to name just one. However, without domain adaptation, they completely fail to detect visible (RGB) video at nighttime or low-resolution images. Figure 1.6 from the website of the YOLOv3 author[†] shows an example of good detection results by YOLOv3 on the RGB image (left). Without domain adaptation, the same detector fails to detect *any* pedestrians at night (the right image).

1.4.1 Multispectral pedestrian detection

Given the advantages of thermal images, recent works on pedestrian detection have investigated the use of thermal sensors as a signal *complementary* to the visible spectrum images. Konig et al. (2017) combined three visible channels and a thermal

[†] `https://pjreddie.com/darknet/yolo/`



(a) Daytime (b) Nighttime

Figure 1.6: Example detection results from the YOLOv3 detector in RGB images. (a) Good results during the daytime. (b) At night, not a single pedestrian is detected.

channel to detect persons in multispectral videos. Many other state-of-the-art multispectral methods, such as those proposed by Wagner et al. (2016); Jingjing et al. (2016), Li et al. (2019), and Li et al. (2018) fused thermal and visible images for multispectral pedestrian detection. Approaches such as these aim to combine thermal and RGB information in order to obtain the most robust possible pedestrian detection at any time of the day. Such detectors require both visible spectrum and thermal images to function. However, this can limit applicability in real applications due to the cost of deploying multiple aligned sensors (thermal and visible). Most importantly, using visible spectrum sensors does not offer the same degree of privacy preservation as using thermal-only images for person detection.

1.4.2 Privacy-preserving person detection

A variety of solutions on how to mitigate privacy problems were discussed in Angelini et al. (2019). For example, using Kinect or other depth sensors for privacy-preservation was considered by Nakashima et al. (2010), while leveraging low resolution video data for privacy-preserving anomaly detection was investigated by Angelini et al. (2019). However, most of the mentioned solutions have limited performance or high installation costs due to reliance on multiple sensors.

Aside from the privacy-preserving affordances offered by thermal imagery, there are also technical and economic reasons to prefer thermal-only detection. Because of this, many recent works do not use visible images but focus only on thermal images for pedestrian detection (John et al., 2015; Herrmann et al., 2018; Baek et al., 2017; Devaguptapu et al., 2019; Guo et al., 2019). They typically yield lower performance than multispectral methods since robust pedestrian detection using thermal-only data is non-trivial, and there is still potential for improvement. In this thesis, we also focus on pedestrian detection using *only thermal imagery*. We focus on how to leverage the advantages of state-of-the-art detection in RGB images and bring them

to the thermal domain via *domain adaptation*.

1.4.3 Domain adaptation

Domain adaptation has a long history for both supervised and unsupervised recognition in computer vision. Domain adaptation attempts to exploit learned knowledge from a source domain in a new target domain. There are many research directions for domain adaptation to the thermal imagery, such as Long et al. (2015) who proposed a Deep Adaptation Network to address domain discrepancy with feature transferability, Masana et al. (2017) who addressed domain transfer and compressibility of deep neural networks. Image-to-image translation using Unified Generative Adversarial Networks (StarGAN) was used to close the gap between two domains by Choi et al. (2017).

Domain adaptation is more challenging with cross-modality data having significantly different distributions (e.g. adapting from the visible to the thermal domain). In recent years, many works has investigated the feasibility of cross-modality adaptation and proposed domain adaptation frameworks to adapt deep learning models from the source modality to target modality (Dou et al., 2018; Chen et al., 2019). Ours is a problem of cross-modality adaptation, however we use *domain adaptation* in a general sense to mean adapting to a new input distribution (i.e. even if this involves a change in modality).

One of the most standard approaches for pedestrian detection in the thermal domain is *transfer learning* which directly fine-tunes a pre-trained model on thermal domain imagery. In this case the adaptation happens through a backpropagation signal coming from the loss at the top of the network down to the beginning of the network where adaptation to new input distribution happens. For this reason, we refer this method of domain adaptation via transfer learning as *top-down domain adaptation*.

This type adaptation, however, results in unacceptable performance. Figure 1.7 shows a detection result using the conventional fine-tuning method and one of our domain adaptation methods (task-conditioned domain adaptation, see Chapter 5). As we can see, top-down domain adaptation (left image) yields questionable performance with one true positive detection and many false-negative detections. On the other hand, with our domain adaptation approach (right image), we can minimize the miss-rate of detection. Because of these considerations, in this these we investigate a range of alternative approaches to adapting RGB pedestrian detectors to the thermal domain.

(a) Top-down domain adaptation (b) Task-conditioned domain adaptation

Figure 1.7: Fine-tuning versus task-conditioned domain adaptation. (a) The standard, top-down approach to domain adaptation using fine-tuning. (b) Our top-down adaptation approach (Chapter 5). Top-down domain adaptation results in a detector that misses many pedestrians, while top-down adaptation results in zero missed and zero false positive detections.

1.5 Contributions of this Thesis

In this thesis we investigate domain adaptation approaches for pedestrian detection in thermal imagery. Our research spans many types of domain adaptation from transfer learning methods to network architecture and data augmentation approach using a generative adversarial network. The contributions of this thesis are:

1. we propose a bottom-up domain adaptation approach based on a simple network segment that adapts thermal inputs to the feature distribution expected by an RGB-trained pedestrian detector;

2. we propose a layer-wise domain adaptation approach and show that, by carefully controlling the *schedule* of network adaptation, state-of-the-art detection results can be obtained using only thermal imagery;

3. we show that thermal detection networks can be *task-conditioned* by training them to simultaneously solve the detection and a related classification task, and then show that such task conditioning significantly improves thermal detection results; and

4. we show how a Generative Adversarial Network can be used to synthesize thermal images during training in order to automatically augment thermal data available for training.

1.6 Organization of this Thesis

The rest of this thesis is organized as follows. In the next chapter, we briefly review related work from the computer vision literature on pedestrian detection and domain adaptation for both multispectral and thermal-only. In chapter 3 we describe three top-down domain adaptation techniques and our proposed bottom-up domain adaptation approach that we apply to the problem of pedestrian detection in thermal imagery. Next, in chapter 4 we describe our layer-wise domain adaptation approach and analyze its performance on two thermal detection datasets. We present our task-conditioned thermal detection network in chapter 5, which we show significantly improves domain adaptation in thermal imagery. Finally, a data augmentation approach using synthetic thermal images generated by GANs for pedestrian detection is described in chapter 6. We conclude in chapter 7 with a discussion of our contribution and future research directions.

Chapter 2

Related Work

In this chapter we review work from the recent computer vision literature most relevant to all of our contributions. In each subsequent chapter we only review the literature most relevant to the content of that chapter.

2.1 Pedestrian Detection in the Visual Spectrum

Pedestrian detection has consistently attracted the attention of the computer vision research community through the years, and the literature review on it is vast (Benenson et al., 2014). With the advent of deep neural networks, higher and higher accuracy has been achieved. Computer vision applications and pedestrian detection have improved significantly in both accuracy and speed, with the detector of Angelova et al. (2015a) a prime example that is accurate enough to be relied upon and is fast enough to run on systems with limited computing power. They proposed a Deep Network Cascade for pedestrian detection that runs at about 15 fps. Detectors based on Convolutional Neural Networks (CNNs) compete for the state-of-the-art on standard benchmark datasets. For example, Tian et al. (2015) used a single task-assistant CNN (TA-CNN) to train multiple tasks from multiple sources to bridge the gaps between different datasets. Their detector is learned by jointly optimizing along with semantic tasks such as pedestrian and scene attribute detection. The speed of the proposed model, however, is limited to 5 fps. Along these lines, the estimation of visibility statuses for multiple pedestrians and recognition of co-existing pedestrians via a mutual visibility deep model was proposed by Ouyang et al. (2016a). Their deep learning model was evaluated on four datasets with an improvement of around 5% - 11% miss rate compared to the SVM method.

A first observation is that Fast/Faster R-CNN have become the predominant architectures for state-of-the-art pedestrian detection because of their flexibility. For example, Li et al. (2017) proposed a Scale-Aware Fast R-CNN model which incorporates a large sub-network and a small sub-network into a unified architecture and

implements a divide and conquer strategy. Their method introduces multiple built-in subnetworks that detect pedestrians with scales from disjoint ranges. With this strategy, each network detects pedestrians of different sizes and then combines the results at the end. Similarly, Zhang et al. (2016) proposed combining a Region Proposal Network (RPN) with Boosted Forests (BF) for pedestrian detection based on Faster R-CNN.

Semantic segmentation has also been shown useful for improving robustness in pedestrian detection via low-level pixel information. For example, Du et al. (2017) integrated a pixel-wise semantic segmentation (SS) network into their deep neural network fusion architecture (F-DNN) as a reinforcement to the pedestrian detection task. Their network fusion architecture first uses a Single-shot Multi-box Detector (SSD) trained as an object detector to generate all possible pedestrian candidates of different sizes. Then, multiple deep neural networks (ResNet-50 and GoogleNet) are used in parallel for further refinement of these pedestrian candidates. They used the SS network, trained on the Cityscapes dataset (Cordts et al., 2016) as a parallel classification network which directly processes input images and produces a binary mask distinguishing pedestrians from background. Finally, they fused the segmentation mask and the original pedestrian candidates using a confidence score. This segmentation network improves pedestrian detection results, however the inference speed is limited at 2.48 seconds per image (around 0.4 fps). In contrast, Brazil et al. (2017a) proposed a framework that fuses a semantic segmentation network into shared feature maps for pedestrian detection. They showed that the additional supervision of the segmentation network is helpful for the downstream pedestrian detector.

Another example is the high-level semantic features for anchor-free pedestrian detection proposed by Liu et al. (2019b). They used ResNet-50 to scan for likely pedestrians in the image. Their detector extracts feature points at different scales and then fuses these multi-scale feature maps into a single one. Finally, a detection head consisting of followed by two prediction layers (one for the central location and the other for the corresponding scale) outputs the detection results. Though this detector reduces miss rate, it operates at only about 3 fps.

The advantage of these methods is that they demonstrate excellent RGB domain performance. However, their applicability under varying illumination conditions is limited. Thus, pedestrian detection is still challenging due to a variety of environmental conditions such as changing illumination, occlusion, and other challenging conditions. This has led the computer vision research community to investigate multispectral methods.

2.2 Multispectral Pedestrian Detection

The rapid advances in deep learning in recent years has led to the development of several detector architectures that have been shown to be general purpose and adaptable to a range of applications. Some examples include Faster-RCNN by Ren et al. (2015), SSD by Liu et al. (2016b), and YOLOv3 by Redmon and Farhadi (2018). Because RGB-based pedestrian detection is limited in challenging conditions such as dark environments, particularly at night or bad weather conditions, there is a growing body of research using thermal imagery alone or in combination with visible imagery for investigating pedestrian detection task. We classify deep learning approaches exploiting the thermal domain into two categories:

- **Multispectral** methods which use both visible and thermal images for training and testing. Such models are typically based on two-stage network architectures with a backbone such as VGG16 (Simonyan and Zisserman, 2014) or Faster-RCNN (Ren et al., 2015).

- **Single-modality methods** which use only thermal or RGB imagery. Single-modality methods from the literature are mainly based on single-pass networks such as SSD (Liu et al., 2016b) or YOLO (Redmon and Farhadi, 2018). During training, single-modality methods can leverage pre-trained models and perform transfer learning from other domains.

The trade-off between these two groups of methods will be discussed in this and the next section.

2.2.1 Multispectral datasets

Several multispectral pedestrian detection datasets, which typically consist of pairs of visible and thermal spectrum images, have been proposed in the literature on multispectral image and understanding. Some notable, ealy datasets include:

- The **LSI Far Infrared Pedestrian** dataset by Olmeda et al. (2013) for tracking and detection consists of 15,224 images of only 164×129 pixels.

- The **CVC-09 Thermal** dataset by Socarras et al. (2013) consists of 5,309 positive images and 2,115 negative images for training, and 5,763 images for testing.

- The **OSU Thermal Pedestrian** dataset by Davis and Keck (2005) contains 284 frames (at 360×240 pixels), multispectral OSU-CT dataset by Davis and Sharma (2007) for object detection in 2007 with around 17,000 color/thermal frames at 320×240 pixels.

Note that these datasets are quite small and contain very low-resolution images captured in video surveillance scenarios. For more details of datasets published before 2014, please see the review by Ghiass et al. (2014). Below we only discuss recently collected datasets containing large numbers of high-resolution images.

The KAIST Multispectral Pedestrian Benchmark. Introduced by Hwang et al. (2015), KAIST is the first well-aligned color-thermal pair dataset for multispectral pedestrian and is the largest thermal-visible dataset up to 2015. The dataset consists of 95,328 aligned visible/thermal image pairs (at 640×480 pixels) taken from a vehicle. The dataset is split into 50,172 images for training and 45,156 for testing. All image pairs are manually annotated with three object classes (person, people, and cyclist) for a total of 103,128 dense bounding box annotations and 1,182 unique pedestrians. For training and evaluation, most published results use the *reasonable* setting by Dollar et al. (2012) and Hwang et al. (2015) which separates daytime and nighttime detection evaluation and excludes heavily occluded and small person instances of fewer than 50 pixels. The final training set contains 19,058 image pairs by sampling every 2 frames from 50,172 image pairs. The test set consists of 2,252 image pairs by sampling every 20 frames from 45,156 image pairs. Because the original annotation of the dataset has many problems, the improved training annotations by Li et al. (2018) and test annotations from Jingjing et al. (2016) are used. Results on the KAIST dataset are evaluated in the daytime and nighttime settings using log average miss rate (LAMR) over false positive per image (FPPI) described by (Dollar et al., 2012).

The CVC-14 Visible-FIR Day-Night Pedestrian Sequence Dataset. Proposed by Gonzalez Alzate et al. (2016) is a multi-modal dataset (far infrared and visible). It is smaller than KAIST, but is also design for benchmarking multispectral pedestrian detection. However, two of the video streams in the CVC-14 dataset are not synchronized (image pairs are not properly aligned). The dataset consists of a total of 8,518 FIR and visible frames recorded during the day and night. Pedestrian bounding boxes larger than 50 pixels were manually annotated.

The FLIR Thermal Dataset was released by the thermal camera manufacturer FLIR (FLIR, 2018). It consists of 10,228 color/thermal image pairs with bounding box annotations for five classes: person, bicycle, car, dog, and other. These images are split into 8,862 for training and 1,366 for testing. However, the color and thermal cameras have different focal lengths and resolutions and are not properly aligned. In order to compare with the state-of-the-art on this dataset, it is important to follow the benchmark procedure described by Devaguptapu et al. (2019). The FLIR dataset aims to enable the community to create the next generation of safer and more efficient ADAS systems using cost-effective thermal cameras.

2.2.2 Multispectral approaches

The combination of thermal and RGB images has been proposed by many works on robust pedestrian detection. For example, Xu et al. (2017) used thermal image features to improve detection results in visible-spectrum images by proposing a cross-modality learning framework including a Region Reconstruction Network (RRN) and Multi-Scale Detection Network (MDN). The RRN is used to learn a non-linear feature mapping between RGB and thermal image pairs. Then, the learned model is transferred to a target domain where thermal inputs are no longer available. The MDN is used for learning an RGB-based pedestrian detector. The advantage of their strategy is that multispectral data are not employed at test time. This is crucial when deploying the application, as only traditional cameras are needed which significantly decreases costs. Moreover, no pedestrian annotations are required in the thermal domain. This greatly reduces human labeling effort and permits the exploitation of large data collections of RGB-thermal image pairs. However, using only the RGB image in the testing phase faces the challenges discussed above. The performance is not comparable with multispectral detectors using both color and thermal data at test time.

As a starting point for multispectral pedestrian detection research, Hwang et al. (2015) proposed the KAIST Multispectral Pedestrian Detection Benchmark, one of the biggest multispectral (thermal-color) datasets for pedestrian detection. They also investigated an extension of aggregated channel features by combining pairs of RGB and thermal images and taking advantage of distinct image channels for improving pedestrian detection and establishing a baseline on the KAIST dataset. Similarly, Gonzalez Alzate et al. (2016) proposed the combination of a patch-based detector using Random Forests on Histograms of Oriented Gradients and Local Binary Patterns for pedestrian detection on the CVC-14 Visible-FIR Day-Night Pedestrian Sequence Dataset Gonzalez Alzate et al. (2016).

In order to better exploit the advantages of thermal imagery, multispectral methods typically use a fusion of visible and thermal features for both training and testing phases. One of the most naive ways is a fusion using two-branch networks, one for visible spectrum input and the other for thermal spectrum input. Many works have investigated several types of fusion for multispectral pedestrian detection. For example, Wagner et al. (2016) proposed a deep CNN model using two types of fusion networks to exploit visible and thermal image pairs. The first is an Early Fusion architecture, which combines the information of two modalities at a pixel level. The second is a Late Fusion CNN, which uses separate sub-networks to generate a feature representation for each branch. An additional fully-connected layer combines these feature representations for the final detection. Similarly, four different network fusion approaches (early, halfway, late, and score fusion) for multispectral pedestrian detection task were proposed by Jingjing et al. (2016). Early fusion concate-

nates the feature maps from color and thermal branches immediately after the first convolutional layer, halfway fusion fuses after the fourth convolutional layer, late fusion fuses by concatenating feature maps at the last fully-connected layer, and score fusion can be thought of as a cascade of two CNNs. Among the four fusion models, halfway fusion achieved the best performance.

Because the combination of thermal and visible images works well in two-stage network architectures, most of the top-performing multispectral pedestrian detection approaches are based on Fast/Faster R-CNN (Ren et al., 2015) using a VGG16 backbone (Simonyan and Zisserman, 2014). For instance, Konig et al. (2017) detected persons in multispectral video using one thermal and three visible channels. They combined a fully convolutional Region Proposal Network (RPN) and a Boosted Decision Trees Classifier (BDT). Starting with individual fusion RPN built upon VGG16 backbones pre-trained on thermal and RGB images, they fused these CNNs halfway through to generate deep multispectral features for the RPN. The proposals are further evaluated using a the BDT to reduce potential false positive detections. Similarly, Faster R-CNN was used by Li et al. (2019) in their Illumination-aware Faster R-CNN method. Two sub-networks with VGG16 backbones take visible and thermal images, respectively, as input and generate proposal bounding boxes. Then, the RPN module computes a fusion weight of the two modalities.

One of the most notable methods is the Multispectral Simultaneous Detection approach of Brazil et al. (2017b). This detector uses a Multispectral Proposal Network (MPN) and a Multispectral Classification Network (MCN). The MPN starts from two separate VGG16 backbones (one for RGB and one for thermal). The two networks are then joined using halfway fusion by concatenating convolutional feature maps before generating candidate bounding boxes. The MCN architecture is designed similarly to the MPN. The total loss of the whole architecture consists of sixteen loss terms, including segmentation, classification, and bounding box loss. With this huge network architecture (and improved training annotations for KAIST), their results represent the multispectral state-of-the-art on the KAIST dataset.

Fritz et al. (2019) used VGG16 as an RPN backbone to analyze generalization ability on three multispectral pedestrian detection datasets. They started from VGG16 networks pre-trained on ImageNet and adapted them to detect in the visible or infrared spectrum. Halfway fusion was then performed to obtain a multispectral model for analyzing generalization ability among chosen datasets. Their experimental results showed that the KAIST Multispectral Pedestrian Benchmark is the best dataset to train well-generalizing multispectral RPNs.

A common assumption in multispectral pedestrian detection is that the color-thermal image pairs are geometrically aligned. However, the modalities are only weakly aligned in practice which degrades pedestrian detection performance in two ways. Firstly, features from different modalities are mismatched in the correspond-

ing positions. Secondly, it can be difficult to cover the objects in both modalities with a single bounding box. Thus, an Aligned Region CNN was proposed by Zhang et al. (2019a) to deal with weakly-aligned multispectral data. A Region Feature Alignment module captures the positional shift and adaptively aligns the region features of the two modalities. Then a multi-modal fusion model called an Aligned Region CNN (AR-CNN) used to performs feature re-weighting that selects more reliable features and suppresses the others.

Several top-performing multispectral pedestrian detectors are built upon anchor-based detectors. Many anchor boxes are needed during training to ensure sufficient overlap with most ground-truth boxes. This causes slowdown during training, and the performance significantly drops when applied to small images. A box-level segmentation learning framework for accurate, real-time multispectral pedestrian detection was proposed by Cao et al. (2019) that eliminates the need for anchor boxes. They took pairs of aligned visible and thermal images and their input bounding box annotations.

A common advantage of these multispectral methods is that there are many ways to fuse image features and improve results. Because they fuse both visible and thermal images to enrich image representation, they usually leverage two-stage frameworks suitable for learning combined representations of two inputs. However, since they are usually based on far more complex network architectures that align modalities at inference time, detection speed usually slows to under 5 fps.

Some approaches emphasize applicability to real-time applications that require efficient pedestrian detection. These utilize one-stage detectors which are typically significantly faster than two-stage architectures. For example, a fast single-pass network architecture (YOLOv2 (Redmon and Farhadi, 2017) with pre-trained weights from the PASCAL VOC 2007) was used by Vandersteegen et al. (2018) for multispectral person detection. The network takes four image channels as input and can detect at about 80 fps. However, the accuracy on the KAIST dataset is limited, especially on nighttime images. To improve nighttime detection, Lee et al. (2018) leveraged a deconvolutional, single-shot multi-box detector (DSSD) proposed by Fu et al. (2017) to exploit the correlation between visible and thermal features. They proposed the deep fusion network taking thermal and visible spectrum images as input. Then, the feature maps are concatenated using halfway fusion before feeding to the DSSD network. Their results on the KAIST dataset showed that the DSSD with a ResNet-101 backbone improves miss rate by about 1% - 2% compared to the traditional SSD network with the VGG16 backbone. Similarly, to balance the trade-off between two-stage detectors which achieve higher accuracy and one-stage detectors that focus on fast performance, two Single Shot Detectors (SSDs) were used by Zheng et al. (2019) in combination with Gated Fusion Units that learn the best combination of feature maps generated by the two SSD branches.

Aside from the advantages of the aforementioned multispectral models to make the most out of both modalities and obtain the state-of-the-art result, there are some disadvantages to multispectral detection:

- Color-thermal image pairs and their annotations might not always be available, as they can be prohibitively expensive to collect and require image alignment to be completely accurate.
- To obtain well-aligned data from two-cameras systems, the sensors must be calibrated and synchronized. Estimating intrinsic and extrinsic parameters in calibration is not an always easy, especially with low-resolution thermal images mounted on a mobile platform. A slight misalignment between color and thermal imagery can reduce the performance of the detector.
- Multispectral models are typically complex and based on two-branch network architectures receiving two inputs. This complexity can result in difficulties with deployment.
- Aside from the technical and economic motivations for preferring thermal-only sensor deployment over multispectral methods, using visible spectrum images does not guarantee the same privacy-preserving affordances offered by thermal-only detectors.

Motivation by these challenges, we focus on improving both performance and speed for pedestrian detection using *only* thermal imagery.

2.3 Single-modality Methods for Thermal Imagery

In part because of the disadvantages listed above, some pedestrian approaches work in single-modality imagery. Most concentrate on how to best leverage the advantages of detectors trained on other modalities while training a detector that does not require these extra modalities at inference time.

2.3.1 Pedestrian detection in thermal imagery

There are a few works that, like ours, focus on pedestrian detection using only thermal imagery. An early example is the work by John et al. (2015), which uses adaptive fuzzy C-means clustering to segment IR images and retrieve candidate pedestrians, then prunes candidate pedestrians by classifying with CNN. The authors report a significant reduction in computational complexity compared to the sliding window approach.

Thermal images are powerful for detecting pedestrians in conditions where color images fail, such as at night. However, during the day other objects in the surroundings are as warm as or warmer than humans, making them less distinguishable. Ghose et al. (2019) addressed the challenge of pedestrian detection in thermal images, especially during the day. They used a Pixel-wise Contextual Attention network (PiCA-Net) to create saliency maps. Then, Faster R-CNN was trained for pedestrian detection using the original thermal image and another channel containing the saliency map.

A few approaches leverage RGB images as data augmentation by performing RGB to thermal image translation. For instance, many data preprocessing steps were applied by Herrmann et al. (2018) to make thermal images look more similar to grayscale-converted RGB images, allowing pre-trained RGB features to be effective in the thermal domain. Then a fine-tuning step was performed on a pre-trained SSD300 detector. Recently, the Cycle-GAN was used by Devaguptapu et al. (2019) as a preprocessing step for image-to-image translation before feeding the input to Faster-RCNN. Their model consists of two branches. One branch is pre-trained on large-scale RGB datasets and fine-tuned using a visual RGB input that obtained using an image-to-image (I2I) translation framework from a given thermal image. The second branch follows the standard training process on a relatively smaller thermal dataset. The multi-modal architecture helps to borrow complex high-level features from the RGB domain to improve object detection in the thermal domain.

The common drawbacks of most of the methods mentioned above are that they use many complex preprocessing steps or use hand-crafted features. As a consequence, their performance suffers, and speed is still low such as Fuzzy C-means by John et al. (2015) requires 2.5 seconds per frame (0.4 FPS) and achieves only 34% miss on KAIST and TPIHOG Baek et al. (2017) needs 40 seconds per frame to reach only a 56.8% miss rate.

2.3.2 Pedestrian detection in thermal imagery at night

One of the biggest advantages of thermal imagery is the ability to detect pedestrians at night. Thus, some works approached the pedestrian detection task with thermal imagery concentrating only on nighttime detection. For example, Heo et al. (2017) used adaptive Boolean map-based saliency (ABMS) to boost the pedestrian from the background based on the particular season. They showed that pedestrians have higher saliency than the background, and the ABMS is used as a hardwired kernel in a saliency feature map combined with YOLOv2 for pedestrian detection. Another work is the nighttime pedestrian detector proposed by Liu et al. (2016a). Firstly, the temperature matrix of infrared images is used to extract candidate pedestrians. Then, the histogram of oriented gradient and intensity (HOGI) feature are extracted from the infrared image. Finally, the HOGI features are employed to train a classifier

based on two kinds of machine learning algorithms. Results show that a pedestrian region of interest (ROI) extraction method based on the temperature of the matrix can greatly overcome the low infrared image resolution, environmental reflection, and stability of the device itself.

The work by Baek et al. (2017) proposed to use Thermal Position Intensity Histogram of Oriented Gradients (TPIHOG) and the Additive Kernel SVM (AKSVM) for nighttime-only detection in thermal imagery. The proposed TPIHOG includes detailed information on gradient location; therefore, it has more distinctive power than the HOG. The AKSVM performed faster and better than the linear SVM in terms of detection performance.

Most of these approaches are designed with two processing steps with handcrafted feature extraction as the first step, limiting their ability to compete with the state-of-the-art. Moreover, focusing on only night-time (discarding day-time) is not a full solution for pedestrian detection applications. We focus on both daytime and nighttime thermal-only detection, but we tackle the problem of transferring knowledge between domains and adapting the learned knowledge from the source domain to the new domain. Our domain adaptation approaches are relatively simple because they are based on the single-stage detector YOLOv3 by Redmon and Farhadi (2018), which can be optimized end-to-end and retains its real-time performance.

2.4 Domain Adaptation

Domain adaptation has a long history for both supervised and unsupervised recognition in computer vision. Domain adaptation attempts to exploit learned knowledge from a source domain in a new target domain. Kouw (2018) discussed transfer learning and domain adaptation and how to generalize from a source to a target domain, including risk minimization and three special cases of dataset shift. Many works have investigated domain adaptation techniques to bridge the gap between domains, such as Long et al. (2015) proposed a Deep Adaptation Network (DAN) architecture to reduce the domain discrepancy by enhancing the feature transferability in order to generalize the domain adaptation scenario. The hidden representations are embedded into a reproducing kernel Hilbert space, where the mean embedding of different domain distributions can be explicitly matched. Another interesting work is Masana et al. (2017) who addressed domain transfer problem with the compression of DNN, which learned representations in the large source domain and exploited on a smaller target domain. They focused on compression algorithms based on low-rank matrix decomposition called Domain Adaptive Low Rank (DALR) method. They analyzed the activation statistics when compressing weights by optimally remove the redundancy in the weights. Their experiments showed that the 6th fully connected layer of VGG19 could be compressed four times

more with only a minor or no loss in accuracy and significantly improved classification results. Their methods allowed for compression down to only 5-20% of the original number of parameters, with only a minor drop in performance for domain-transferred networks.

For domain adaptation between thermal and visible domain, an early work in thermal infrared person detection is that of Herrmann et al. (2018), which uses domain adaptation based on feature transformation techniques (inversion, equalization, and histogram stretching) to transform thermal images as close as possible to the color. The translation problem can be seen as two kinds of learning: (1) the supervised learning problem where the network needs to access corresponding pairs of instances from both domains; and (2) the unsupervised learning problem, where no such paired instances are available. In order to focus on the latter case, which is more difficult but at the same time more realistic as acquiring the dataset of paired images is often difficult in practice, a deep architecture called an Invertible Autoencoder (InvAuto) was proposed by Teng et al. (2018). It treats an encoder as an inverted version of a decoder in order to decrease the trainable parameters of image translation processing. Similarly, Wang et al. (2018c) proposed a model including a CycleGAN (Zhu et al., 2017b) to translate thermal facial images into visible images, and a detector with Pix2Pix to locate important facial landmarks on visible faces and help the generative network to generate more realistic images that are easier to be recognized. Their experiments demonstrated that the faces generated have good visual quality and maintain identity preserving features.

One of the closest methods to one of our is the work of Devaguptapu et al. (2019) who proposed a "pseudo-multimodal" object detector trained on natural visible image data to improve object detection performance in thermal images. Firstly, they used the image-to-image translation framework Cycle-GAN (Zhu et al., 2017b) to automatically generate pseudo-RGB equivalents of a given thermal image. Then, a Multi-modal Thermal Object Detection Methodology (MMTOD) for object detection in the thermal image, which consists of two branches of the network, one for the thermal image input and the other for the RGB input. The positive signal of their method is they do not need the pairs of image training example for two modalities, for each thermal image input, they used image-to-image (I2I) translation network to generate a pseudo-RGB, then two these inputs are passed through a detector which is Faster R-CNN with Region Proposal Network. Their experimental result on two datasets FLIR and KAIST dataset outperformed the baseline. However, the result is still limited compared to state-of-the-art methods, and the speed is only 0.11 second per image (around 9 FPS).

2.5 Contributions of this Thesis with Respect to the State-of-the-art

In all of our approaches, instead of relying on multispectral input we focus on thermal-only detection using a single-pass detector which is both fast and accurate during the day and at night. Moreover, we do not approach the thermal domain by learning cross feature representations between visible and thermal. Instead, we approach the detection problem by preserving the learned feature representations from the source RGB domain via a variety of novel adaptation strategies. Our thermal pedestrian detectors are, to the best of our knowledge, the state-of-the-art in single-modality pedestrian detection today. Several of our approaches are even competitive with multispectral detectors despite the fact that we use no visible spectrum imagery at detection time.

Chapter 3

Bottom-up Domain Adaptation for Pedestrian Detection in Thermal Imagery[†]

In this chapter, we investigate two domain adaptation techniques for fine-tuning a YOLOv3 detector to perform accurate and robust pedestrian detection using *thermal* images. Our approaches are motivated by the fact that thermal imagery is *privacy-preserving* in the sense that person identification is difficult or impossible in low-resolution images. Results on the KAIST dataset show that our approaches perform comparably to state-of-the-art approaches and outperform the state-of-the-art on nighttime pedestrian detection, even outperforming multimodal techniques that use both thermal and visible spectrum imagery at test time.

3.1 Introduction

Object detection is a classical problem in computer vision, and person and pedestrian detection is one of the most important topics for safety and security applications such as video surveillance, autonomous driving, person re-identification, and numerous others. The estimate of the total number of installed video surveillance cameras is significantly increasing. The advent of autonomous driving promises to add many more cameras, all detecting and observing humans in public spaces.

Recent works on pedestrian detection have investigated the use of thermal imaging sensors as a complementary technology for visible spectrum images (Vandersteegen et al., 2018). Approaches such as these aim to combine thermal and RGB image information in order to obtain the most robust possible pedestrian and per-

[†]Portions of this chapter were published in: M. Kieu, A. D. Bagdanov, M. Bertini, A. Del Bimbo, "Domain Adaptation for Privacy-preserving Pedestrian Detection in Thermal Imagery." *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*, 2019.

son detection and any time of the day or night. Such detectors require both visible spectrum and thermal images to function.

Citizens are naturally concerned that being observed violates their right to privacy. In this chapter we are interested in investigating the limits of pedestrian detection using thermal imagery alone. The advantages of thermal imagery in this respect is shown in figure 1.5, illustrating how thermal images can retain distinctive image features for pedestrian detection while preserving privacy. Our hypothesis is that thermal images can guarantee the balance between security and privacy concerns.

The rest of this chapter is organized as follows. In the next section, we briefly review related work from the computer vision literature on pedestrian detection, domain adaptation, and thermal imaging. In section 3.3 we describe conventional top-down domain adaptation approaches for pedestrian detection problem in thermal imagery. Then, in section 3.4 we detail our proposed bottom-up approach to domain adaptation that we apply to the problem of privacy-preserving person detection. We report on a range of experiments conducted in section 3.5, and conclude in section 3.6 with a discussion of our contributions.

3.2 Related Work

In this section we review some recent work related to pedestrian detection, domain adaptation, and computer vision for thermal imagery.

Person and pedestrian detection. The literature, both classical and contemporary, on pedestrian detection is vast (Benenson Rodrigo and Bernt, 2014). With the advent of deep neural networks in recent years, pedestrian detection is achieving higher and higher accuracy (Angelova et al., 2015b). However, pedestrian detection remains a challenging task due to occlusion, changing illumination and variation of viewpoint and background (Ouyang et al., 2016b). Several CNN-based pedestrian detection methods compete for the state-of-the-art on standard benchmark datasets for pedestrian detection as described in section 2. Examples include Pedestrian Detection aided by Deep Learning Semantic Tasks (Yonglong Tian and Tang, 2014), Scale-Aware Fast RCNN (Li et al., 2015), Learning Mutual Visibility Relationship (Ouyang et al., 2016b). These state-of-the-art techniques use RGB images as input, while our goal is to investigate the potential of detection in thermal imagery alone.

Domain adaptation. Domain adaptation has played a main role in both supervised and unsupervised recognition in computer vision. Domain adaptation attempts to exploit learned knowledge from the source domain in the target domain. One of our approaches was inspired by the AdapterNet (Hazan et al., 2018), which proposed adding a new shallow Convolutional Neural Network (CNN) before the original model that transforms the input image the target domain before passing through an

unmodified network trained in the source domain. Several works have tried to mitigate the distance between the two domains by applying transformation techniques. For example, the idea from Herrmann et al. (2018) was to transform infrared data (thermal domain) as close as possible to the color domain by using feature transformations: inversion, equalization and histogram stretching. A deep architecture, called Invertible Autoencoder (InvAuto), introduced a method to treat an encoder as an inverted version of a decoder in order to decrease the trainable parameters of image translation processing (Teng et al., 2018).

Pedestrian detection exploiting thermal imagery. Several works demonstrate that using thermal images in combination with RGB images can improve object detection results. An example is the work by Xu et al. (2017), which suggests a method based on a cross-modality learning framework focusing only on visible images at test time. During training time, they use thermal image features to boost visible detection results. Their method has two main phases: Region Reconstruction Network (RRN), for learning a non-linear feature mapping between visible and thermal image pairs, and a Multi-Scale Detection Network (MDN) which performs pedestrian detection from visible images by exploiting the cross-modal representations learned with RRN.

A variety of recent works leverage two-stage network architectures to investigate the combination of visible and thermal features. Wagner et al. (2016) investigated two types of fusion networks. Another approach is the ACF+T+HOG technique (Jingjing et al., 2016) which considers four different network fusion approaches (early, halfway, late, and score fusion). Konig et al. (2017) introduced a combination Fully Convolutional Region Proposal Networks (RPN) and Boosted Decision Trees Classifier (BDT) for person detection in multispectral video. Illumination-aware Faster R-CNN (IAF RCNN) (Li et al., 2019) and Illuminating Pedestrians via Simultaneous Detection and Segmentation (Brazil et al., 2017b) used the Faster R-CNN detector to perform pedestrian detection on paired RGB and thermal imagery. A Fusion architecture network (MSDS-RCNN) including a multispectral proposal network (MPN) and a multispectral classification network (MCN) was proposed by Chengyang Li and Tang (2018). This fusion network currently yields the best results on both visible and thermal image pairs on the KAIST dataset.

In a slightly different direction, the combination of HOG and SVM proposed by Baek et al. (2017) focused on only nighttime detection. Their method uses a Thermal Position Intensity Histogram of oriented gradient (TPIHOG) and the additive kernel SVM (AKSVM) for training and testing.

Differing from most of the above works which used two-stage detectors, some papers utilize a one-stage detector (Lee et al., 2018; Vandersteegen et al., 2018). The authors of (Lee et al., 2018) used a deconvolutional single shot multi-box detector (DSSD) to exploit correlation between visible and thermal features for person detec-

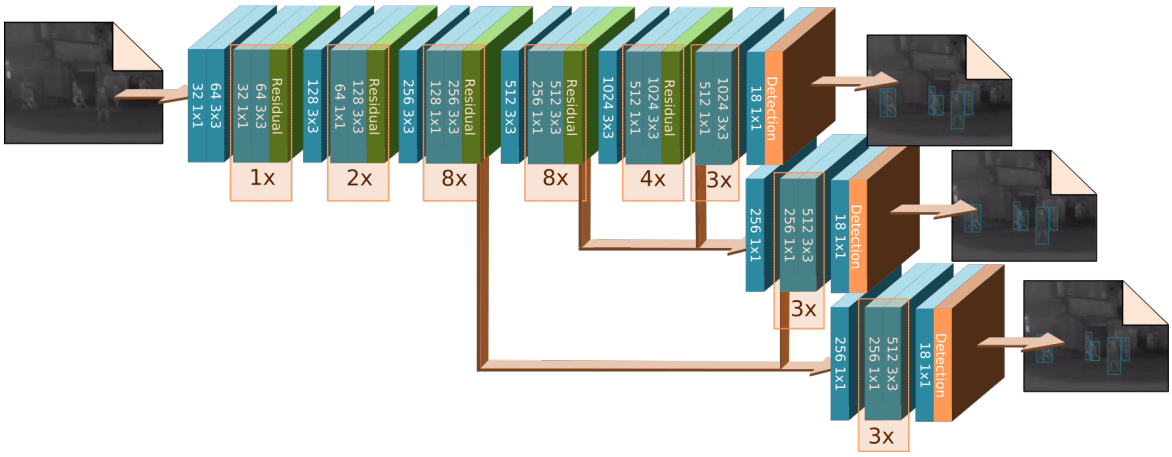


Figure 3.1: The YOLOv3 architecture. $k\times$ indicates the repetition of blocks k times.

tion. A fast RGB single-pass network architecture (YOLOv2 (Redmon and Farhadi, 2017)) was adopted by Vandersteegen et al. (2018) for fine-tuning for person detection.

In this chapter, we investigate the potential of two domain adaptation approaches including top-down domain adaptation and bottom-up domain adaptation for pedestrian detection tasks in thermal-only domain. Our extensive experimental results show that our approaches outperform the state-of-the-art both single modality and multimodal approaches in night-time on the challenge KAIST dataset.

3.3 Top-down Domain Adaptation Approaches

In this section we describe the approaches to domain adaptation that we will later evaluate in section 3.5. All of our approaches use the YOLOv3 detector which is adapted to a target domain through a sequence of domain adaptation steps.

One of the most standard approaches to adaptation of deep models to new domains is *fine-tuning*. Since fine-tuning typically works by decapitating the original network and training via backpropagation from the top of the network down to the bottom, we refer to the use of fine-tuning for domain adaptation as *top-down adaptation*. The top-down adaptation approaches we consider are based on transfer learning and use one of the fastest and most accurate detectors available today: YOLOv3 by Redmon and Farhadi (2018), which is pre-trained on ImageNet and subsequently fine-tuned on the MS COCO dataset by Lin et al. (2014). We adapt the YOLOv3 detector to the new target domain through a sequence of domain adaptation steps. YOLOv3 is a very deep detection network with 106 layers and three detection heads for detecting objects at different scales as illustrated in figure 3.1. YOLOv3 uses a fully-convolutional residual network as its backbone. The network is coarsely struc-

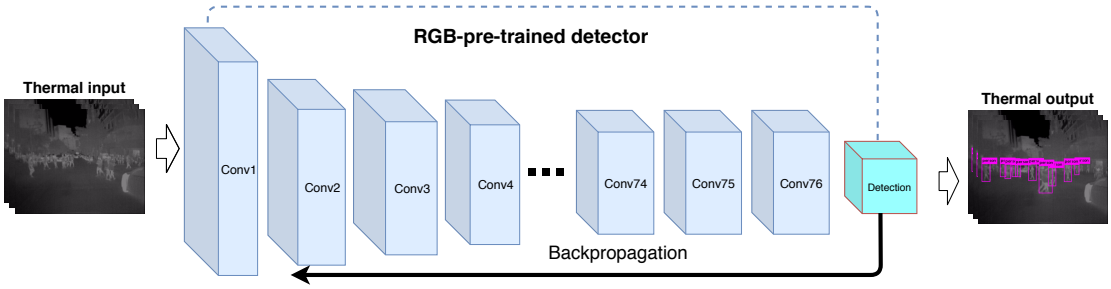


Figure 3.2: Top-down domain adaptation refers to using fine-tuning to adapt a detector to a new domain (e.g. thermal imagery). Adaptation to the new input distribution happens only via back-propagated loss from the end of the network (at the top) down to the new input distribution.

tured into five residual *groups*, each consisting of one or more residual *blocks*. As we see in figure 3.1, these five *groups* include 23 residual *blocks*, each consisting of two-convolutional layers with residual connections adding the input of each block to the output.

We refer to this as *top-down adaptation* because of the way fine-tuning on the new domain happens only via back-propagation where the supervision signal comes from the loss at the *end* (i.e. the *top* of the network), down to the new input distribution. In figure 3.2 we illustrate this top-down adaptation, which refers to the fine-tuning approach to the new input distribution (thermal domain in our case). We fine-tune the pre-trained RGB detector to adapt to the new thermal input.

In the descriptions below, we use a notational convention to refer to each technique that indicates which image modalities are used for training and testing. For example, the technique reported as TD(VT, T) is Top-Down domain adaptation, with adaptation on Visible spectrum images, followed by adaptation on Thermal images, and finally tested on Thermal images. The three top-down domain adaptation approaches we consider are:

- **Top-down visible: TD(V, V)**: This domain adaptation approach directly fine-tunes YOLOv3 on visible images in the target domain for pedestrian detection. Testing was performed on visible spectrum images. This experiment mainly served as the baseline for comparison with single modality techniques on visible imagery.

- **Top-down thermal: TD(T, T)**: This approach directly fine-tunes YOLOv3 on only thermal images by duplicating the thermal image three times, once for each input channel of the RGB-trained detector. Testing was performed only on thermal imagery (no RGB images are available at test time). This exper-

iment served as the baseline for the comparison with single modality techniques and domain adaptation in thermal imagery alone.

- **Top-down visible/thermal: TD(VT, T)**: This approach is a variant of the two top-down approaches described above. First, we adapt YOLOv3 to the visible spectrum pedestrian detection domain, and then we fine-tune that detector on thermal imagery. Testing was performed only on thermal images (no RGB images available). The idea here was to determine if knowledge from the visible spectrum could be retained and exploited after the final adaptation to the thermal domain.

3.4 Bottom-up Domain Adaptation: BU(VAT, T)

A hypothesis of ours is that in top-down domain adaptation, as described in the previous section, early convolutional layers are difficult and slow to adapt to the new input distribution due to their distance from the backpropagated loss. Here we propose a type of *bottom-up* domain adaptation which first trains a bottom-up adapter segment and then proceeds to fine-tune the detector using a top-down loss. A conceptual schema of this approach is given in figure 3.3. The main components of our bottom-up domain adaptation approach are as follows.

3.4.1 Notation

Let $f_\Theta(\mathbf{x})$ represent the detector (YOLOv3 in our case) parameterized by parameters $\Theta = \{\theta_1, \theta_2, \ldots, \theta_N\}$, where θ_i represents the parameters of the ith layer of the network. We use the notation $\Theta_{n:m}$ to denote the parameters of layers n through m of the network f (so $\Theta = \Theta_{1:N}$). Similarly, we use $f_{n:m}$ to represent the forward pass of the network f from layer n through layer m. We assume f to be pre-trained for detection in the RGB domain – in our experiments we start from the network TD(V, V) described above.

3.4.2 Adapter segment training

The first step in bottom-up domain adaptation is to train an *adapter segment* to mimic RGB feature activations when given *thermal* images as input. We create a new network segment f' identical to the first m layers of f (i.e. copying the weights of TD(V, V). Given paired visible/thermal spectrum images $(\mathbf{x}_v, \mathbf{x}_t)$ (such as those available in KAIST), we train the parameters θ' of f' using original detection loss for pedestrian detection with thermal images as input. The main idea of the adapter segment is to intervene an early stage of the RGB-trained detector network and to train this adapter segment to adapt to the thermal domain. The number of layers of adapter

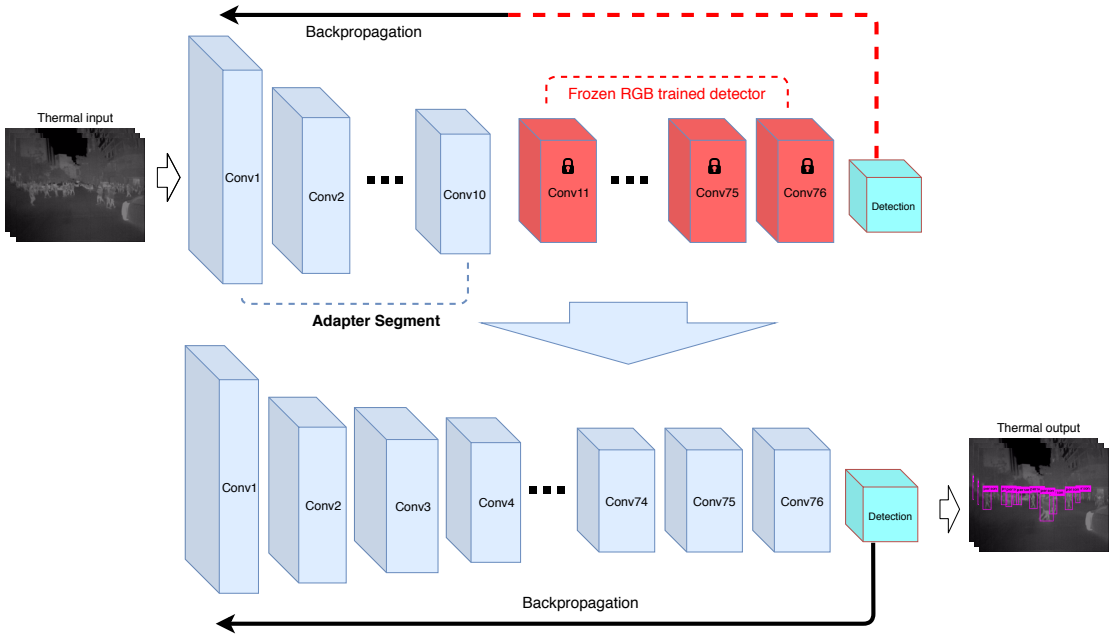


Figure 3.3: **Bottom-up domain adaptation**. Starting from an RGB-trained detector, an *adapter segment* is trained to take thermal images as input and produce convolutional features similar to the features of the original network. When the adapter segment training has converged we reconnect the adapter segment to the *RGB-trained detector* for the final fine-tuning.

segment m is a hyperparameter, and in early experiments we found $m = 10$ to be a good point of intervention for adapter segment training.

As illustrated in the top of the figure 3.3, the main idea of the Adapter Segment is to intervene at some early stage of the RGB-trained detector network and to train a parallel branch that takes only thermal imagery as input and *matches* as best as possible the RGB feature maps at the point of intervention. In our implementation, we decapitate the YOLOv3 network after the first ten convolutional layers and train a ten-layer adapter segment to match the RGB-network using only thermal images as input.

The starting point for this approach is the TD(V, V) network described above. That is, the detector weights we start from are already adapted to the KAIST domain on visible images. We then train the *adapter segment* to adapt to the thermal image input from the KAIST training set. Thus, VAT indicates adapting on *visible domain* (V) first, then training the *adapter segment* (A) on thermal domain, and finally reconnecting the adapter segment for the final fine-tuning on the *thermal domain* (T). The adapter segment is the "A" in the "VAT" mnemonic: BU(VAT, T).

3.4.3 Fine-tuning of entire detector

After adapter segment training has converged, we reconnect the newly trained adapter segment to the original RGB-trained detector for the final fine-tuning of the whole detector on thermal images (as illustrated at the bottom of figure 3.3). To do this we train the final detection network $f_{m+1:N}(f'(\mathbf{x}_t))$ using only thermal images $\mathbf{x}_t$ and the original loss of the YOLOv3 network.

3.5 Experimental Results

In this section we report results of experiments we performed to evaluate the performance of adapted detectors for pedestrian detection in thermal imagery. More importantly, we detail the dataset, evaluation metrics which will be applied in the next chapters.

To evaluate our proposed approaches to domain adaptation we used a standard benchmark dataset of RGB/thermal image pairs and standard evaluation prototols.

3.5.1 The KAIST Multispectral Pedestrian Detection Benchmark

All experiments of this thesis were performed on the publicly available KAIST Multispectral Pedestrian Detection Benchmark (Hwang et al., 2015). KAIST is the only large-scale dataset with well-aligned visible/thermal pairs (Devaguptapu et al., 2019), and it contains videos captured both during the day and at night. The dataset consists of 95,328 aligned visible-thermal image pairs in total split into 50,172 for training and 45,156 for testing. The dataset contains 103,128 annotations of 1,182 unique pedestrians. There are several reasons we choose KAIST dataset as a main dataset for our experiments, for example: (1) KAIST is one of the largest available well-aligned visible/thermal image pairs dataset for pedestrian detection task; (2) over the past four years, it attracted a vast of research and the annotations of the KAIST dataset have been improved for both training and test set (Jingjing et al., 2016; Li et al., 2018);

According to the official sampling method from the baseline (Hwang et al., 2015) and the some recent papers (Konig et al., 2017; Vandersteegen et al., 2018), we sampled images to obtain the train set and test set, the sampled procedure will be explained again as following:

- For test set, we sampled every 20 frames from 45,156 images (this means we get 1 frame for each 20 frames). Finally, we obtained 2,252 visible/thermal image pairs for the test set, of which 797 pairs are captured at night and 1,455 pairs at day time.

Figure 3.4: Example thermal/RGB image pairs from the KAIST dataset.

- For the train set, we also sampled every 2 frames and filtering (e.g occlusion, the bounding box under 50 pixels) from 50,172 images to obtain 19,058 visible/thermal pairs.

The train and test annotation was provided publicly by the baseline (Hwang et al., 2015). Figure 3.4 gives some example thermal/visible image pairs from the KAIST dataset (Hwang et al., 2015).

3.5.2 Evaluation metrics

For evaluating our proposed methods on KAIST and to compare with the state-of-the-art, we strictly follow the metrics and the *reasonable* setting provided by the KAIST benchmark (Hwang et al., 2015) and the state-of-the-art results (Li et al., 2018; Vandersteegen et al., 2018; Jingjing et al., 2016; Konig et al., 2017).

We use standard evaluation metrics for object detection, namely log average miss rate as a function of False Positives Per Image (FPPI). The log-average miss rate is calculated for thresholds in the range of $[10^{-2}, 10^{0}]$ with an Intersection over Union (IoU) under the *reasonable* setting (Dollar et al., 2012; Hwang et al., 2015; Jingjing et al., 2016). The *reasonable* setting is composed of *day-time*, *night-time*, and *all* (*both day and night time*) sets of images. For computing miss rates (MR), an Intersection over Union (IoU) threshold of 0.5 is used to calculate True Positive (TP), False Positives (FP) and False Negatives (FN). True Positive (TP) is counted if a detected bounding box is matched to a ground-truth box with an Intersection of Union (IoU) of 50% or greater. Unmatched detected and ground truth boxes are considered False Positives and False Negatives, respectively. The MR is computed by averaging miss rate (false negative rate) at nine False Positives Per Image (FPPI) rates evenly spaced

in log-space. The evaluation source code we used from the work by Vandersteegen et al. (2018), which is an updated version of the Matlab code from Dollar et al. (2012).

3.5.3 Implementation details

We used the YOLOv3 (Redmon and Farhadi, 2018) detector to evaluate our approach on KAIST. Our detectors were implemented using PyTorch, and we trained every domain adaptation strategy for 50 epochs with a learning rate 0.0001 and the Adam optimizer.

3.5.4 Comparative performance analysis

The plots in figure 3.5 show detailed results for our approach and those described by Vandersteegen et al. (2018) in terms of precision/recall (left column) and log-average miss rate (right column). The plots also break down results in terms of time-of-day: first row averaged over all times, second row daytime only, third row nighttime only.

From the results in figure 3.5 we can make several observations. First of all, for combined day and night results (first row) multimodal techniques like YOLO_TLV which exploit both thermal and visible spectrum images at test time are superior to our domain adaptation approaches which use only thermal imagery. Surprisingly, however, the gap between bottom-up domain adaptation BU(VAT, V) and YOLO_-TLV is only about 4% in log-average miss rate, which is quite promising.

The reason that multimodal approaches outperform domain adaptation seems to be due to the advantage they have when detecting during the day. In the second row of figure 3.5, in fact, we see that the technique exploiting visible spectrum images during at test time on daytime images outperform all our approaches which only use thermal imagery.

Or two domain adaptation approaches, both top-down and bottom-up, outperform all other techniques when testing at nighttime only (third row of figure 3.5). Though this is not very surprising, of particular note is the fact that performing domain adaptation on to *visible* images before adapting to thermal input only is beneficial. This can be seen in the difference between TD(VT, T), BU(VAT, T) – both of which start by fine-tuning YOLOv3 on KAIST visible images – and TD(T, T), which directly fine-tunes YOLOv3 on thermal images. This seems to indicate that both top-down and bottom-up domain adaptation are able to retain and exploit some domain knowledge acquired when training the detector on visible spectrum imagery.

As a final comment, we note that the BU(VAT, T) approach requires significantly less training time than the others. In only 15 epochs it converged to 84.4% precision, which is the same result for top-down adaptation after 50 epochs. Bottom-up

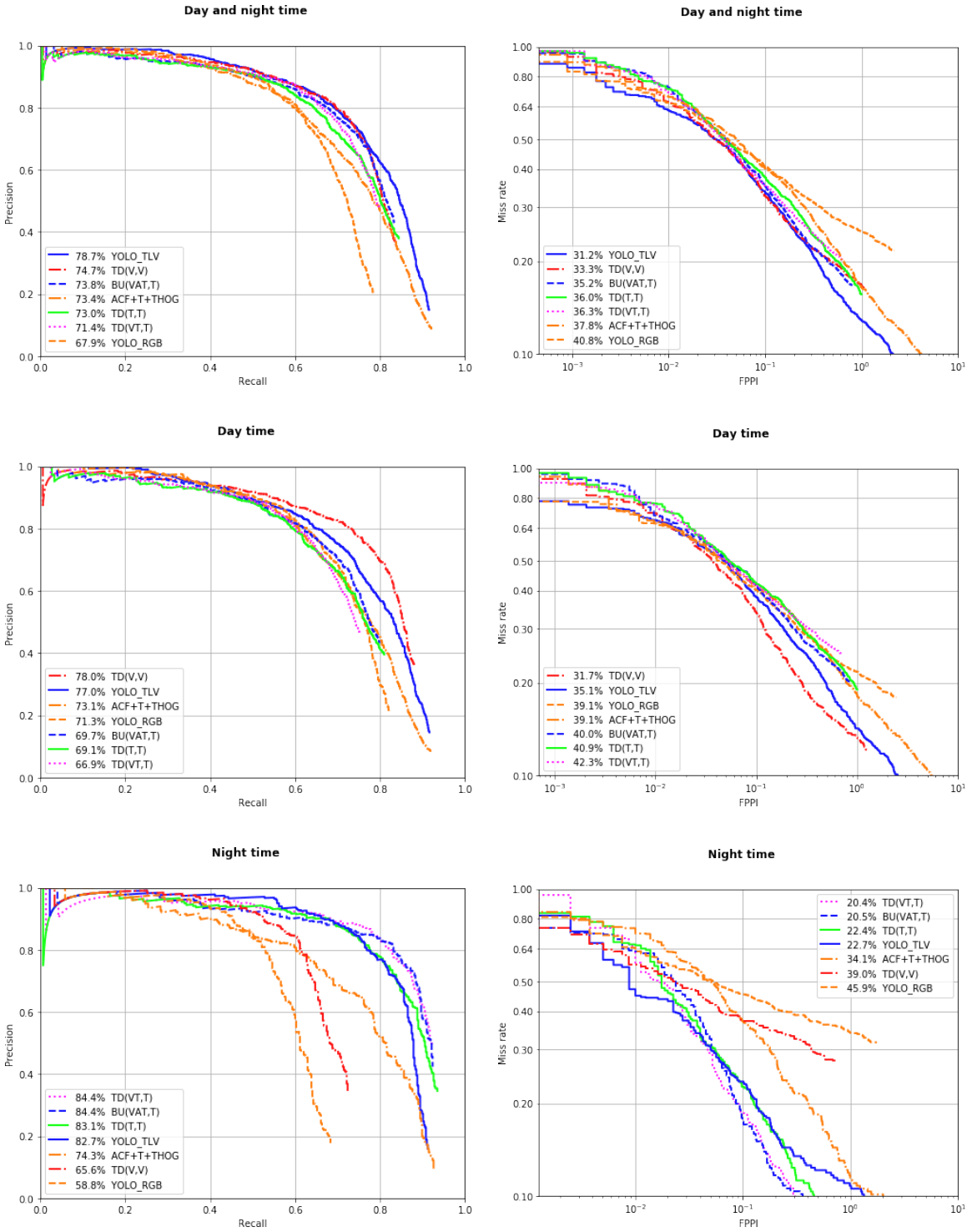


Figure 3.5: **Comparative performance analysis**. Precision/Recall (left, higher is better) and Log-average Miss Rate (right, lower is better) of our method and other state-of-the-art papers are given. See text for detailed analysis.

Table 3.1: **Log-average Miss Rate (%) on KAIST dataset (lower is better)**. The final two columns (V:Visible, T:Thermal) indicate which image modality is used at *test time*. Our approaches outperform all single-modality techniques from the literature, and outperform all methods at night.

Method	MR all	MR day	MR night	V	T
KAIST baseline (Hwang et al., 2015)	64.76	64.17	63.99	✓	✓
Late Fusion (Wagner et al., 2016)	43.80	46.15	37.00	✓	✓
Halfway Fusion (Jingjing et al., 2016)	36.99	36.84	35.49	✓	✓
RPN+BDT (Konig et al., 2017)	29.83	30.51	27.62	✓	✓
IATDNN+IAMSS (Guan et al., 2018)	**26.37**	**27.29**	24.41	✓	✓
YOLO_TLV (Vandersteegen et al., 2018)	31.20	35.10	22.70	✓	✓
DSSD-HC (Lee et al., 2018)	34.32	-	-	✓	✓
RRN+MDN (Xu et al., 2017)	49.55	47.3	54.78	✓	
TPIHOG (Baek et al., 2017)	-	-	57.38		✓
SSD300 (Herrmann et al., 2018)	69.81	-	-		✓
Ours: TD(V,V)	33.30	31.70	39.00	✓	
Ours: TD(T,T)	36.00	40.90	22.40		✓
Ours: TD(VT,T)	36.30	42.30	**20.40**		✓
Ours: BU(VAT,T)	35.20	40.00	20.50		✓

adaptation seems to be an effective way to accelerate top-down adaptation through fine-tuning.

In table 3.1 we provide a comparison of our methods and 10 others methods from the state-of-the-art. Our approaches outperform all other single modality techniques (both visible- and thermal-only). Compared to multi-model approaches, we outperform all of them at nighttime, and comparably on all.

3.5.5 Qualitative evaluation

In figure 3.6 we show some example detection results on the KAIST dataset for our BU(VAT, T) domain adaptation approach in daytime (first row) and nighttime (second row). Note how, even though person identification is impossible in all of the example images, the detector adapted using bottom-up domain adaptation is able to detect pedestrians even in the presence of occlusion, scale variation, and changing illumination conditions.

3.6 Conclusions

In this chapter we investigated the potential of two domain adaptation strategies for adapting pedestrian detectors to work in the thermal domain. The goal of this work is to achieve the best possible person detection performance while relying *solely* on

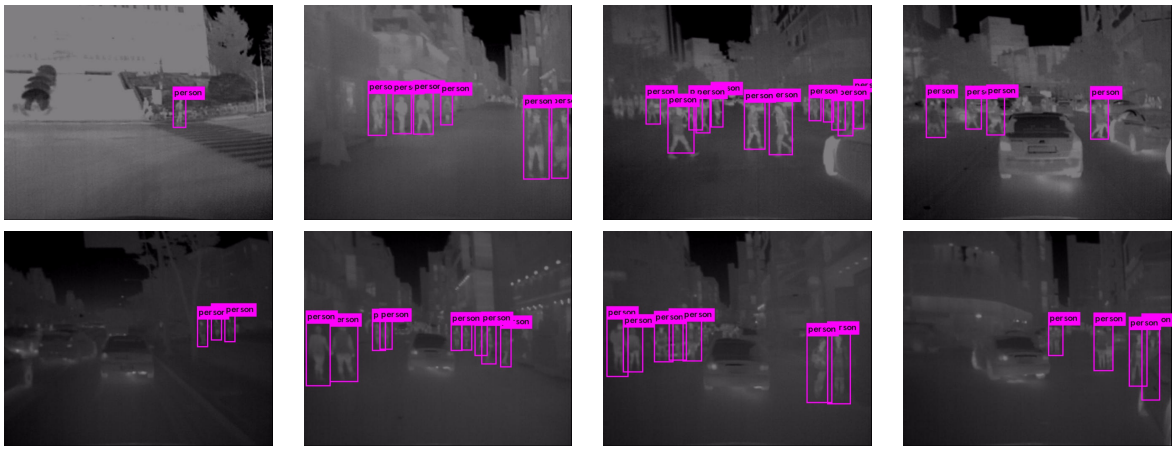


Figure 3.6: **Qualitative results on the KAIST test set**. The first row gives example detections on daytime images from KAIST, and second row on nighttime images. Even in the presence of occlusions and scale variations, thermal imagery retains enough information to effectively perform pedestrian detection – day or night – in a privacy-preserving way without using any visible imagery at detection time.

thermal spectrum imagery. This is motivated by the *privacy-preserving* aspects of thermal images, since persons are difficult, if not impossible, to reliably identify in thermal images.

Our results indicate that relatively simple domain adaptation schemes can be effective, and that the resulting detectors can outperform multimodal approaches (i.e. those that use thermal *and* visible images at test time) at nighttime, and can perform comparably when testing on day night images combined. Moreover, results seem to indicate that a first adaptation to visible imagery can be useful to acquire domain knowledge that can then be exploited after final adaptation to thermal domain.

Chapter 4

Layer-wise Domain Adaptation for Pedestrian Detection in Thermal Images[†]

Building on our earlier work on bottom-up domain adaptation (see the previous Chapter) for privacy-preserving pedestrian detection, we proposed a new type of bottom-up domain adaptation strategy, which we call *layer-wise domain adaptation*. We conducted an extensive experimental evaluation comparing top-down and bottom-up domain adaptation, as well as layer-wise adaptation on more datasets. Our layer-wise domain adaptation approach also includes two steps: first, training an adapter segment corresponding to initial layers of the RGB-trained detector adapts to the new input distribution; then, we reconnect the adapter segment to the original RGB-trained detector for final adaptation with a top-down loss, the main difference is to train the adapter segment. To the best of our knowledge, our layer-wise domain adaptation approach outperforms the best-performing single-modality pedestrian detection results on KAIST, and outperforms the state-of-the-art on FLIR©.

4.1 Introduction

State-of-the-art work on pedestrian detection has mostly concentrated on using multispectral images combining visible and thermal images for training and testing such as (Wagner et al., 2016; Konig et al., 2017; Guan et al., 2018; Li et al., 2018, 2019). Only a few single-modality detection works focus only on thermal images (John et al., 2015; Herrmann et al., 2018; Baek et al., 2017; Devaguptapu et al., 2019; Kieu et al., 2019). Robust pedestrian detection on only thermal data is a non-trivial task

[†]Portions of this chapter were published in: M. Kieu, A. D. Bagdanov, M. Bertini, "Bottom-up and Layer-wise Domain Adaptation for Pedestrian Detection in Thermal Images", *ACM Transactions on Multimedia*, 2020.

and there is still a large potential to improve the thermal detection performance. Our previous work probed the limits of pedestrian detection using thermal imagery alone (Kieu et al., 2019). We investigated three top-down domain adaptation approaches and proposed a bottom-up domain adaptation approach, which outperform state-of-the-art pedestrian detection at nighttime in thermal imagery. Exploiting only thermal data is a fundamental advantage of our work (visible data are not employed at both training adaptation and the test phase), this is crucial when deploying surveillance systems under a variety of environmental conditions.

In addition to extending our work on bottom-up domain adaptation for thermal pedestrian detection, in this chapter we focus on improving pedestrian detection results on thermal-only data by proposing a new, layer-wise domain adaptation strategy which gradually adapts early convolutional layers of a pre-trained detector to the thermal domain. The main motivation for layer-wise adaptation is that, from previous work on bottom-up adaptation, we recognized that adaptation of *early* layers of the network helped the most to preserve the learned knowledge from the visible domain which is useful for the adaptation to the thermal domain. Our hypothesis is that if we adapt slowly from the bottom up in a layer-wise manner, it should improve adaptation. Through extensive experimental evaluation on two datasets and an analysis and interpretation of the contribution of bottom-up adaptation we show that, though only exploiting thermal imagery at test time, our domain adaptation approaches outperform state-of-the-art thermal detectors on the KAIST Multispectral Pedestrian Dataset (Hwang et al., 2015) and FLIR Starter Thermal Dataset (FLIR, 2018). Moreover, our bottom-up and layer-wise adaptation approaches outperform many state-of-the-art *multispectral* detection approaches which exploit both thermal and visible spectra at test time.

The contributions of this work are:

- We propose a new type of bottom-up domain adaptation which adapts the pre-trained detector in a *layer-wise* manner. The result shows that the relatively simple bottom-up strategy better preserves learned features from the visible domain and lead to robust pedestrian detection results in the final detector.

- We give a detailed comparison between three top-down domain adaptation approaches and our two proposed bottom-up adaptation approaches. Our experiments show that our bottom-up and layer-wise adaptation approaches consistently outperform top-down adaptation.

- To the best of our knowledge, we obtain the best detection result on the FLIR© dataset (FLIR, 2018), and we are also the best detection result on the KAIST dataset (Hwang et al., 2015) compared to all existing single modality approaches. Moreover, by exploiting only thermal imagery on KAIST dataset, we outper-

form many the state-of-the-art multispectral pedestrian detectors. which use both visible and thermal for training and testing.

The rest of this chapter is organized as follows. In the next section, we describe our proposed approach to domain adaptation that we apply to the problem of pedestrian detection in thermal imagery. In section 4.3, we report on a range of experiments conducted, and section 4.4 compares our result with state-of-the-art results. In section 4.5, we conclude with discussion of our contribution and future research directions.

4.2 Layer-wise Domain Adaptation

This section describes our proposed approaches, which build upon our earlier work on domain adaptation approaches for pedestrian detection in chapter 3. The goal of our earlier domain adaptation methods showed that relatively simple domain adaptation on only thermal image can outperform many state-of-the-art approaches. Here we introduce a simple layer-wise technique that significantly boosts performance of pedestrian detection in thermal imagery.

We hypothesize that the fine-tuning slowly from the bottom of the network should preserve more knowledge from the original domain. Here we propose a new type of bottom-up domain adaptation, which we call *layer-wise* domain adaptation. It progressively fine-tunes each layer of the network starting from the *bottom* of the network. Layer-wise adaptation also includes two stages: first, layer-wise adaptation to adapt slowly to the new input distribution. Then, a final fine-tuning phase that trains the whole pipeline with a top-down loss. The conceptual schema of our layer-wise domain adaptation approach is given in figure 4.1.

Layer-wise domain adaptation proceeds as follows:

1. **Layer-wise adaptation**: We start from the TD(V, V) network described in section 3.3 – i.e. from an RGB-trained detector already adapted to the new domain in the visible spectrum. We gradually train the initial layers of the YOLOv3 network using thermal images from the training set. As illustrated in the upper part of figure 4.1, the main idea is to adjust the RGB-trained detector network to adapt slowly to the thermal input from bottom of the network up to the top. At epoch i we freeze parameters from the layer $3i + 1$ to N while fine-tuning. That is, at each epoch another three layers are unfrozen until the entire network is being fine-tuned. In the experiments we denote this approach as BU(VLT, T), the "L" in the "VLT" signifies training with layer-wise adaptation.

2. **Fine-tuning of the entire detector**: After adapter segment training has converged, the entire detector trained using end-to-end fine-tuning shown at the

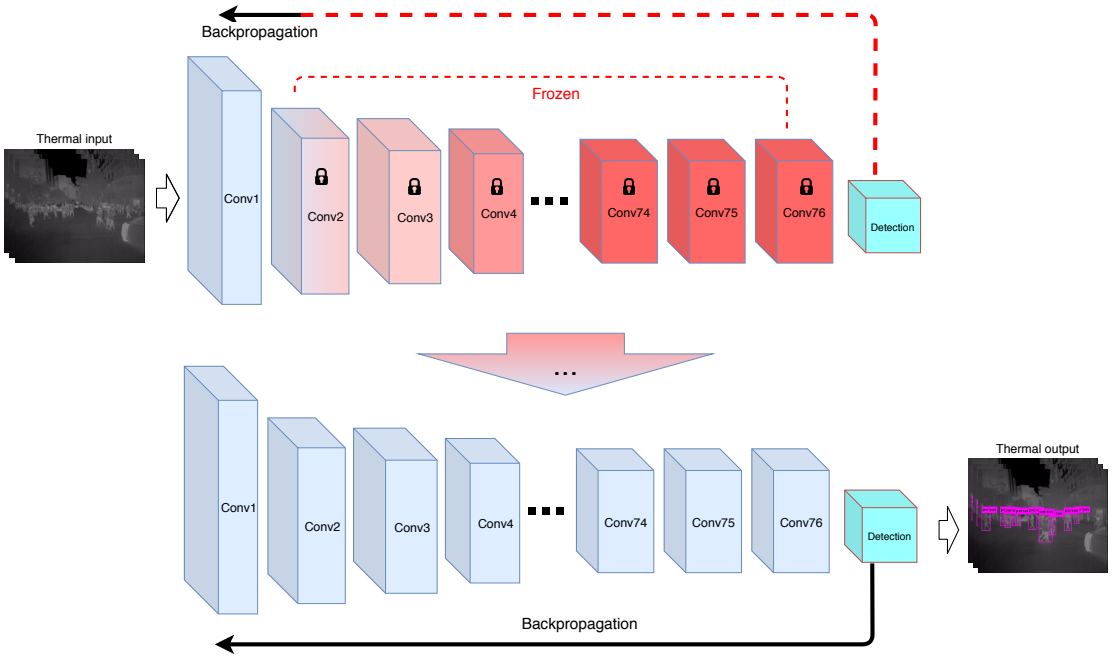


Figure 4.1: **Layer-wise domain adaptation**. Instead of adapter segment training, in layer-wise adaptation we *gradually* adapt layers during fine-tuning. This is done by freezing layers during training and progressively unfreezing them. After gradually including all layers in the training a final fine-tuning of the entire detector pipeline on thermal images is performed.

bottom of figure 4.1. Note that no RGB images are used and the whole pipeline is trained using only thermal images.

We also experimented with a variant of layer-wise adaptation that is more similar to the bottom-up adaptation strategy described by Kieu et al. (2019). Instead of gradually unfreezing layers during adaptation, we freeze only the parameters from the layer 11 to layer N for the first 50 epochs, and then unfreeze them for the remaining 50 epochs to fine-tune the entire network. An overview of this bottom-up strategy is given in figure 4.1. We refer to this bottom-up variant as BU(VAT, T) in the experimental results.

4.3 Experimental Results

In this section we report on a number of experiments we performed to evaluate our domain adaptation approaches with respect to the state-of-the-art in single- and multi-modal detection.

4.3.1 Datasets

All our approaches are evaluated and compared to the state-of-the-art on two public datasets: the KAIST multispectral pedestrian benchmark (Hwang et al., 2015) which described in section 3.5.1 and the FLIR Starter Thermal Dataset (FLIR, 2018). We chose this dataset because of its consistent annotation and its use in other work (Devaguptapu et al., 2019). More specifically, the two datasets used in our experimental evaluation are:

The KAIST dataset. by Hwang et al. (2015) was described detail in the section 3.5.1. However, in this experiment, we used 2 different things compared with the previous chapter in section 3.5.1 as following:

- For the train set, we follow other methods (Li et al., 2018) and other papers, we sample images every 2 frames from training videos and exclude heavily occluded instances and small instances under 50 pixels (height of pedestrian < 50 pixels). The final training set contains **7,601** training images. Noted that, the number of images of train set is less than in previous chapter cause the filtering procedure.
- Because the original annotation of KAIST dataset had problematic (Jingjing et al., 2016). We used the improved annotation of both train set from Li et al. (2018) and test set from Jingjing et al. (2016).

This modify version of KAIST dataset including the number of train set and improved annotation will be used in this chapter and next chapters.

The FLIR dataset by FLIR (2018) was released by the thermal camera manufacturer FLIR.© It consists of 10,228 color/thermal image pairs with bounding box annotations for five classes: person, bicycle, car, dog, and other. These images are split into 8,862 for training and 1,366 for testing. However, the color and thermal cameras have different focal lengths and resolutions and are not properly aligned. In order to compare with the state-of-the-art on this dataset, we follow the benchmark procedure described by Devaguptapu et al. (2019). We evaluate only on thermal images and three object categories: person (28,151 instances), car (46,692 instances), and bicycle (4,457 instances). Some example images from the FLIR Starter Thermal Dataset are given in figure 4.2.

4.3.2 Evaluation metrics

For evaluation, we strictly follow the *reasonable* setting provided by the KAIST benchmark (Hwang et al., 2015). We used the standard precision and log-average miss rate (MR) evaluation metrics for measuring object detection as defined by Dollar et al. (2012). To calculate MR a True Positive (TP) is counted if a detected bounding

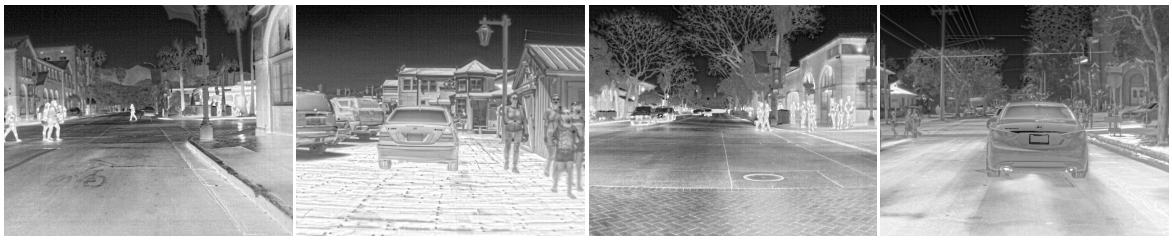


Figure 4.2: Examples from FLIR Starter Thermal Dataset

box is matched to a ground-truth box with an Intersection of Union (IoU) of 50% or greater. Unmatched detected and ground truth boxes are considered False Positives and False Negatives, respectively. The MR is computed by averaging miss rate (false negative rate) at nine False Positives Per Image (FPPI) rates evenly spaced in log-space.

For consistent comparison with the state-of-the-art, we use mean Average Precision (mAP) on the FLIR dataset, while on the KAIST dataset we plot the MR over false positive per images and precision over recall curves to compare to the state-of-the-art. The results from our previous paper (Kieu et al., 2019) trained on the old annotation and setting from the KAIST baseline (Hwang et al., 2015). In this paper, we re-experiment all of our methods on the new training annotation, which provided by Li et al. (2018). The result is significantly improved and comparable with the state-of-the-art. This confirms the critical role of annotation for training the deep neural network.

4.3.3 Experimental setup

All of our models were implemented in PyTorch and source code and pretrained networks are available.* Rather than set apart a fixed validation set, at each epoch we set aside 10% of the training images to use for validation at that epoch. We use mini-batch stochastic gradient descent (SGD) with momentum of 0.9 and do not use the learning rate warm-up strategy of original YOLO model. Training begins with a learning rate of 0.001, and when the training loss no longer decreases and the validation recall no longer improves we decrease the rate by a factor of 10. Training is halted after decreasing the learning rate twice in this way. All models were trained on a GTX 1080 for a maximum of 50 epochs with a batch size of 8. We keep all hyperparameters the same as the original settings of the YOLOv3 model (Redmon and Farhadi, 2018).

*https://github.com/mrkieumy/YOLOv3_PyTorch

Table 4.1: Comparison with state-of-the-art **single-modality approaches** in term of **log-average Miss Rate on KAIST dataset (lower is better)**. Our approaches outperform all others in all conditions (day/night/all).

Detector	all	day	night	test images
KAIST_RGB (Li et al., 2018)	45.70	36.00	68.30	RGB
RRN+MDN (Xu et al., 2017)	49.55	47.30	54.78	RGB
KAIST_thermal (Li et al., 2018)	35.70	40.40	25.20	thermal
TPIHOG (Baek et al., 2017)	-	-	57.38	thermal
SSD300 (Herrmann et al., 2018)	69.81	-	-	thermal
Saliency Maps (Ghose et al., 2019)	-	30.40	21.00	thermal
Bottom-up (Kieu et al., 2019)	35.20	40.00	20.50	thermal
Ours: TD(V, V)	34.75	**29.77**	46.25	RGB
Ours: TD(T, T)	31.10	37.30	16.70	thermal
Ours: TD(VT, T)	30.67	37.42	15.45	thermal
Ours: BU(VAT, T)	26.26	32.84	11.95	thermal
Ours: BU(VLT, T)	**25.61**	32.69	**10.87**	thermal

4.4 Comparison with the State-of-the-art

4.4.1 Performance on KAIST

The KAIST multispectral pedestrian benchmark is a challenging dataset with both nighttime and daytime video described in section 4.3.1. We divide our comparison between single-modality detectors and multi-modal detectors from the literature. Noted that in this chapter, we re-experiment all of our previous methods in chapter 3 on the new training annotation, which described in section 4.3.1. The result is significantly improved and comparable with the state-of-the-art. This confirms the critical role of annotation for training the deep neural network.

Comparison with other single-modality methods. Table 4.1 compares the performance of our approaches with state-of-the-art single-modality approaches (i.e. using only thermal or visible imagery) in terms of miss rate (MR). Our approaches outperform all existing single-modality methods by a large margin in all conditions (day, night, and all). Our layer-wise adaptation obtains the best result with 25.61% MR at all and 10.87% MR at nighttime, improving on the current state-of-the-art by 9.59% at all and 9.63% at night. Our BU(VAT, T) approach reaches 26.26% combined day/night MR and 11.95% MR at nighttime and obtains the second best result compared to the best state-of-the-art of 35.2% MR at all and 20.50% MR at night. Note that we exploit thermal imagery alone for training domain adaptation, while some of the state-of-the-art single-modality methods exploit both color and thermal at training time (Xu et al., 2017; Guan et al., 2018). Among existing single-modality approaches, our detectors are the best on the KAIST dataset in all conditions.

Comparison with both single- and multi-modality approaches. Table 4.2 compares our approaches and more than fourteen other single- and multi-modal from the literature. The last two columns indicate the type of imagery used at test time. The first group contains results for multispectral detectors using both visible and thermal imagery for training and testing. The second group contains single-modality approaches, and our results are in the last group.

We draw a number of conclusions from these results. First of all, from the MR combined results (all), we note that multimodal techniques like MSDS (Li et al., 2018) or IATDNN+IAMSS (Guan et al., 2018) are superior to our domain adaptation approaches. This seems to be due to the advantage they have when detecting during the day and thus exploiting visible imagery. Secondly, our domain adaptation approaches, both top-down and bottom-up, outperform all other single-modality techniques and many multimodal techniques. Thirdly, looking at the nighttime results, our bottom-up domain adaptation BU(VLT, V) is the best result with 10.87% MR. This surpasses the all state-of-the-art approaches in both single- and multi-modal detection. This demonstrates the potential of our domain adaptation methods to capture useful information from RGB detectors and adapt them to nighttime.

Of particular note is the fact that performing domain adaptation on visible images before adapting to thermal input is beneficial. This can be seen in the difference between BU(VAT, T) and BU(VLT, T) – both of which start by fine-tuning TD(V, V) on KAIST visible images – and TD(T, T), which directly fine-tunes YOLOv3 on thermal images. This seems to indicate that both bottom-up domain adaptation approaches can retain and exploit domain knowledge acquired when training the detector on visible spectrum imagery. Notably, slow layer-wise adaptation, BU(VLT, T), helps robust pedestrian detection at night and outperforms other bottom-up methods.

The plots in figure 4.3 provide a more detailed picture of our approaches and the state-of-the-art in terms of precision/recall (left column) and log-average miss rate (right column). The plots also break down results in terms of time-of-day: the first row averaged over day and night, the second row daytime only, and the third row nighttime only. The Intersection of Union (IoU) used is the standard 0.5. The results in the plot are slightly different those originally published because: (1) the authors of Li et al. (2018) said that the number in their official article is calculated by the average of 5 runs; and (2) Vandersteegen et al. (2018) used the overlap IoU 0.4 because they said YOLO had trouble with small objects. All results are generated by their detector result files using the framework provided by Vandersteegen et al. (2018). The results reported in the original papers are given in table 4.2.

The results from figure 4.3 show that the ranking is similar to that reported in table 4.2. We are in the top three results during the day and combined (all). We are also the best results at night. Note that the MSDS result plotted here is higher

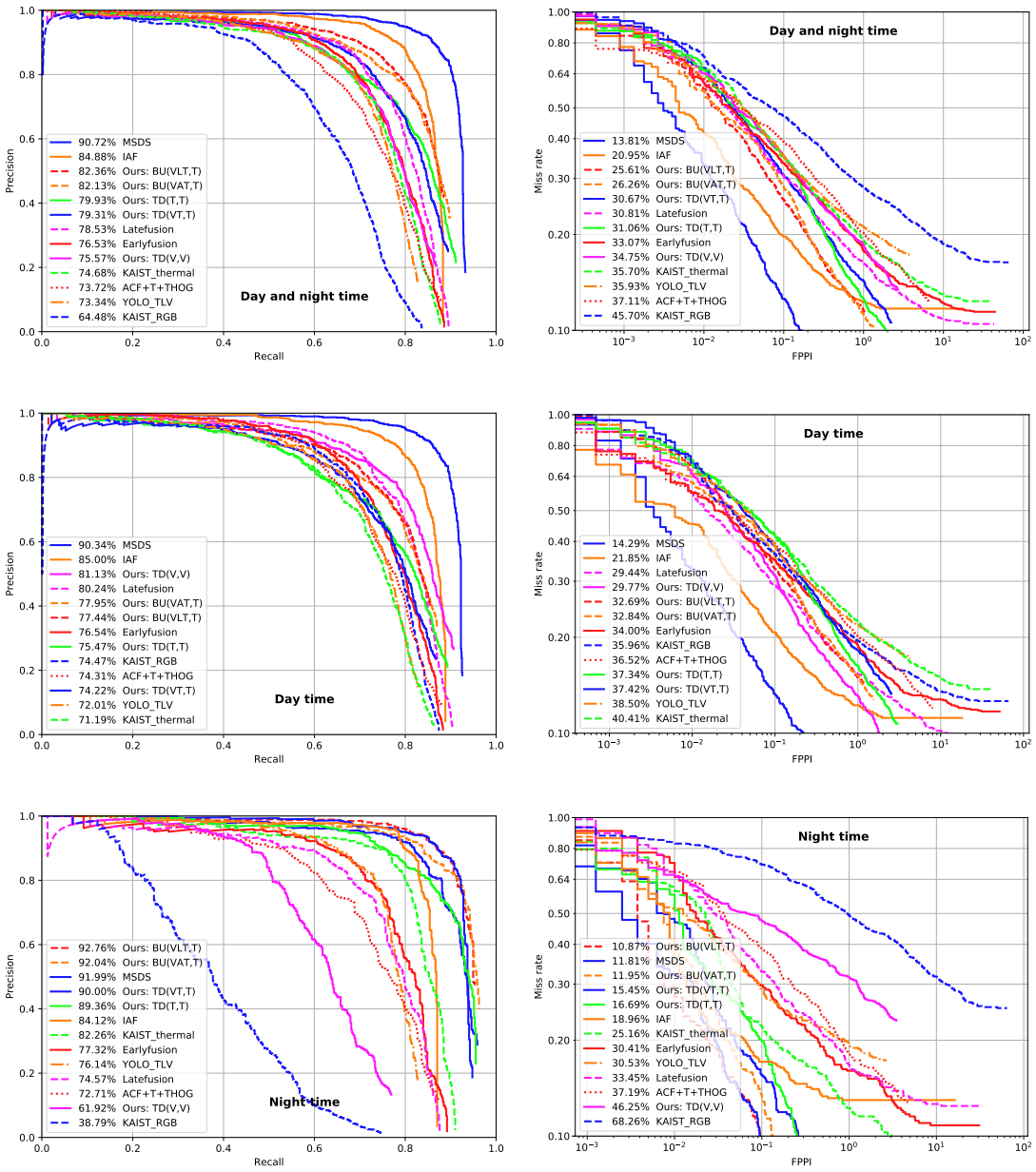


Figure 4.3: **Comparative performance analysis**. Precision/Recall (left, higher is better) and Log-average Miss Rate (right, lower is better) of our method and other state-of-the-art papers are given. See text for detailed analysis.

Table 4.2: **Log-average Miss Rate on KAIST dataset (lower is better)**. The final two columns (V:Visible, T:Thermal) indicate which image modality is used at *test time*. Our approaches outperform all single-modality techniques from the literature, and outperform all methods at night.

Method	MR all	MR day	MR night	V	T
ACF+T+THOG (Hwang et al., 2015)	64.76	64.17	63.99	✓	✓
Latefusion (Wagner et al., 2016)	43.80	46.15	37.00	✓	✓
Halfwayfusion (Jingjing et al., 2016)	36.99	36.84	35.49	✓	✓
RPN+BDT (Konig et al., 2017)	29.83	30.51	27.62	✓	✓
IATDNN+IAMSS (Guan et al., 2018)	26.37	27.29	24.41	✓	✓
IAF (Li et al., 2019)	15.73	14.55	18.26	✓	✓
MSDS (Li et al., 2018)	**11.63**	**10.60**	13.73	✓	✓
YOLO_TLV (Vandersteegen et al., 2018)	31.20	35.10	22.70	✓	✓
DSSD-HC (Lee et al., 2018)	34.32	-	-	✓	✓
GFD-SSD (Zheng et al., 2019)	28.00	25.80	30.03	✓	✓
RRN+MDN (Xu et al., 2017)	49.55	47.3	54.78	✓	
KAIST_RGB (Li et al., 2018)	45.70	36.00	68.30	✓	
TPIHOG (Baek et al., 2017)	-	-	57.38		✓
SSD300 (Herrmann et al., 2018)	69.81	-	-		✓
KAIST_thermal (Li et al., 2018)	35.70	40.40	25.20		✓
Bottom-up (Kieu et al., 2019)	35.20	40.00	20.50		✓
Ours: TD(V, V)	34.75	29.77	46.25	✓	
Ours: TD(T, T)	31.06	37.34	16.69		✓
Ours: TD(VT, T)	30.67	37.42	15.45		✓
Ours: BU(VAT, T)	26.26	32.84	11.95		✓
Ours: BU(VLT, T)	25.61	32.69	**10.87**		✓

Table 4.3: Comparative performance analysis on the FLIR dataset.

Method	Bicycle	Person	Car	mAP
Baseline	39.7	54.7	67.6	54.0
MMTOD-UNIT (Devaguptapu et al., 2019)	49.4	64.5	70.8	61.5
Our: TD(T,T)	51.9	75.5	86.9	71.4
Our: BU(AT,T)	56.1	**76.1**	**87.0**	73.1
Our: BU(LT,T)	**57.4**	75.6	86.5	**73.2**

than numbers in their published paper. Importantly, our domain adaptation approaches, both top-down and bottom-up adaptation, surpass all other methods at nighttime and are comparable with the other two multispectral methods for combined day/night (all).

4.4.2 Performance on FLIR

Table 4.3 compares our approaches with state-of-the-art on the FLIR Starter Thermal Dataset (FLIR, 2018). Results on this dataset are measured using average precision (AP) for each class and the mean Average Precision (mAP) over all classes. Note that the FLIR dataset has five categories, but the baseline and the state-of-the-art results reported only three of these: person, car and bicycle. From these results we see that our approaches, both top-down and bottom-up, outperform the baseline and the state-of-the-art on all classes and in overall mAP. Our layer-wise adaptation is the best result with 73.2% mAP, improving on the current state-of-the-art by 11.7% mAP. Our bottom-up approach BU(AT,T) also obtains 87.0% precision on cars, advancing the current state-of-the-art by 16.2% average precision.

4.4.3 Qualitative analysis of detector adaptation

In figure 4.4 we plot the average gradient magnitudes of every layer of the network during the first epoch of fine-tuning for three different adaptation methods: top-down, bottom-up and layer-wise. These plots show the capacity of the methods to adapt network weights during adaptation to the new domain. We see from these plots that the gradient magnitudes for layer-wise adaptation are highest for all layers – especially for the *early* convolutional layers where the network must adapt the most to the new input domain. The gradient magnitudes for bottom-up adaptation are also larger than simple fine-tuning (top-down adaptation), which illustrates the positive effect the adapter network has on domain adaptation.

In order to better understand how layer-wise adaptation improves internal feature representations for detection in thermal images we used the Gradient weighted Class Activation Map (Grad-CAM) (Selvaraju et al., 2017) visualization technique on input images from the KAIST and FLIR datasets for all three adaptation approaches (top-down, bottom-up, and layer-wise). These visualizations are shown in figure 4.5. Grad-CAM heatmaps were computed at layer 52 of the adapted YOLOv3 using the backpropagated loss from the medium-scale detection head. We see in these visualizations that the layer-wise network has learned to attend to more areas of the image salient to pedestrian detection compared to the bottom-up and top-down adapted networks. This explains how layer-wise adaptation leads to more correct positive detections on average.

Figure 4.6 shows detection results on three images by two methods. The first row gives results of top-down adaptation (TD(VT,T)), and the second row results of bottom-up adaptation (BU(VAT,T)) on the same images from the KAIST dataset.

As we can see on the figure 4.6, *bottom-up adaptation* results in more True Positive bounding box detections than *top-down* (in the first and the second images). Similarly, *Top-down adaptation* results in more False Positive detections than the *bottom-up*

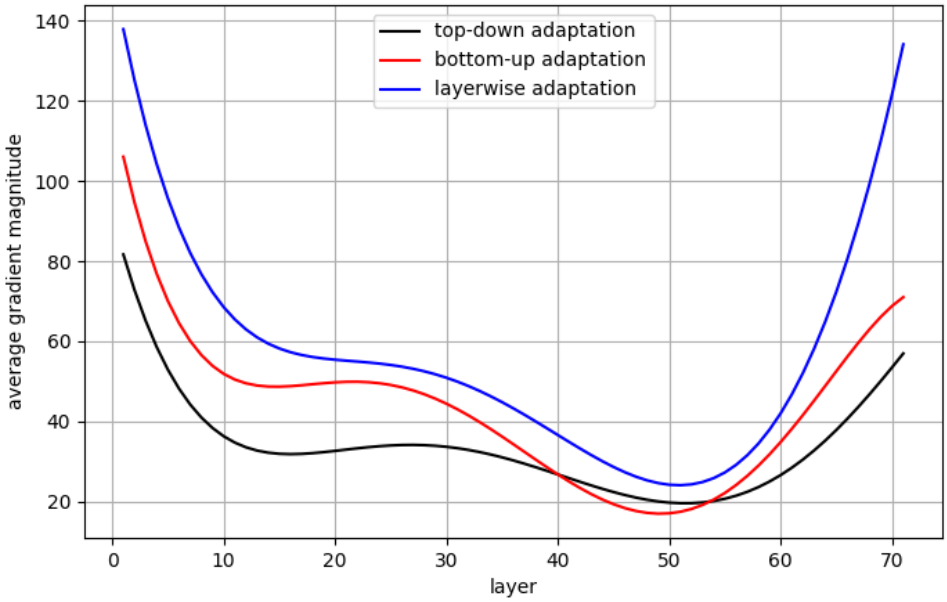


Figure 4.4: Average gradient magnitudes per layer for each adaptation method during first epoch of fine-tuning.

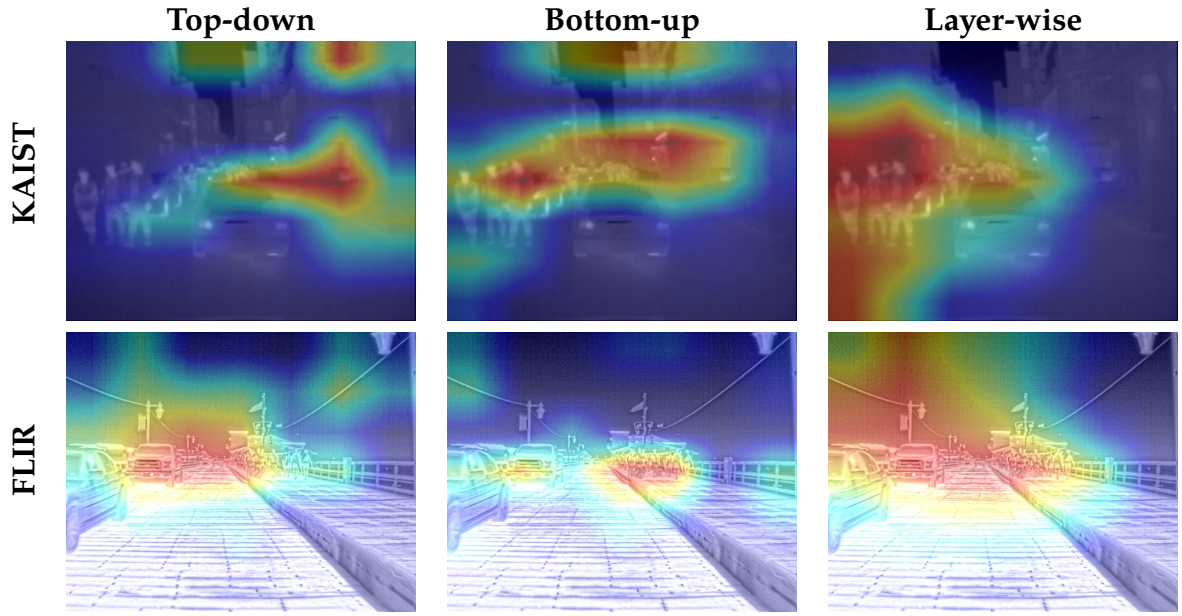


Figure 4.5: Grad-CAM visualization of feature importance for three adaptation methods. Red areas indicate which parts of the image contribute most to the features used in the detection heads of the adapted networks. Note how feature importance for layer-wise adaptation is spread across more pedestrians compared to the other two adaptation methods.

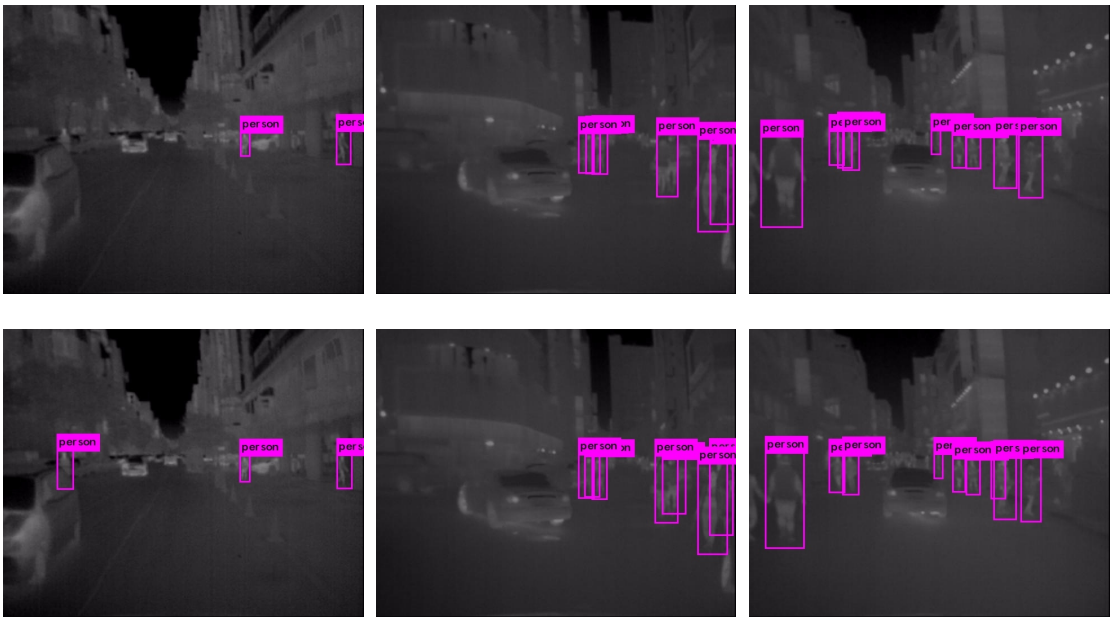


Figure 4.6: The first row gives three frames of detections resulting from top-down domain adaptation. The second row gives results of bottom-up adaptation on the same frames. Even though person identification is extremely difficult low-resolution images, thermal imagery can retain distinctive image features to effectively perform pedestrian detection, and our *bottom-up adaptation* made more True Positive and less False Positive detection result than *top-down adaptation.*

adaptation on the last image. This is consistent with with our experimental results that show *bottom-up adaptation* is superior to *top-down adaptation*. We believe this is because bottom-up adaptation preserves more knowledge which has been learned from the visible domain before adapting to the thermal domain. Moreover, we note that the *bottom-up adaptation* approach requires significantly less training time than *top-down adaptation*. In only 15 epochs it converges to about 82.13% precision, which is a higher than top-down adaptation after 50 epochs. *Bottom-up adaptation* seems to be an effective way to accelerate *top-down adaptation* through fine-tuning. Last but not least, even though person identification is extremely difficult in low-resolution thermal images, thermal imagery retains enough information to effectively perform pedestrian detection in a privacy-preserving way without using any visible spectrum imagery at detection time.

Looking at our results on KAIST in the last row of table 4.1 and the table 4.2, we make some observations:

- Firstly, the technique exploiting visible spectrum images during the day outperforms all our approaches which only use thermal imagery. This is expected as the daytime images are similar to visible images than thermal images.

- The methods TD(VT, T), BU(VAT, T), and BU(VLT, T), which start from TD(V, V), surpass TD(T, T). This shows that the first adaptation on the visible domain helps the network adapt better to the final, thermal target domain. This opens an opportunity to leverage transfer learning from other datasets for robust pedestrian detection.
- Our best thermal detection networks work extremely well at night, but only modestly well on daytime imagery. We believe that daytime and nighttime detection require different features and filters. Thus, there is still plenty of opportunity for improvement in thermal-only detection results.
- Moreover, as we can see our two bottom-up results, in tables 4.1, 4.2, and 4.3, the layer-wise adaptation BU(VLT, T) and BU(VAT, T) are always superior to top-down methods on both datasets. This seems to indicate that a slowly adaptation from the bottom of the network better preserves visible feature, which helps maintain robust detection at night.

4.5 Conclusions

In this chapter, we described the potential of the bottom-up domain adaptation approach for pedestrian detection tasks, and we also proposed an effective layer-wise domain adaptation strategy for pedestrian detection in thermal imagery. The goal of our research is to close the performance gap between pedestrian detection exploiting only thermal imagery and multispectral approaches using both visible and thermal images for training and testing.

The results on two datasets show that our relatively simple domain adaptation schemes are effective, and our results outperform all state-of-the-art single-modality methods on two datasets. Exploiting only on thermal domain, our detectors perform comparably with the state-of-the-art and outperform many multispectral approaches on KAIST. Furthermore, the results reveal that a preliminary adaptation to visible spectrum images is useful to acquire domain knowledge that can be exploited after the final adaptation to the thermal domain. As far as we know, ours result is the best result in thermal imagery on FLIR datasets.

Chapter 5

Task-conditioned Domain Adaptation in Thermal Imagery[†]

In the previous two chapters, we showed how bottom-up and layer-wise domain adaptation help preserve the knowledge from the visible domain when performing domain adaptation to the thermal domain. The results at nighttime are extremely good, however, there is still a large gap between detector performance on RGB and thermal imagery during the day. In this chapter we propose a novel approach to domain adaptation that significantly improves pedestrian detection performance in the thermal domain. The key idea behind our technique is to adapt an RGB-trained detection network to simultaneously solve two related tasks. An auxiliary classification task that distinguishes between daytime and nighttime thermal images is added to the main detection task during domain adaptation. The internal representation learned to perform this classification task is used to condition a YOLOv3 detector at multiple points in order to improve its adaptation to the thermal domain. We validate the effectiveness of task-conditioned domain adaptation by comparing with the state-of-the-art on the KAIST Multispectral Pedestrian Detection Benchmark. To the best of our knowledge, our proposed task-conditioned approach achieves the best single-modality detection results.

5.1 Introduction

Pedestrian detection problem is particularly challenging in many common contexts such as limited illumination (nighttime) or adverse weather conditions (fog, rain, dust) (Li et al., 2018; Kieu et al., 2020a). For these reasons, detectors exploiting thermal imagery have been proposed as suitable for robust pedestrian detection (Kieu

[†]Portions of this chapter were published in: My Kieu, Andrew D. Bagdanov, Marco Bertini, Alberto Del Bimbo, "Task-conditioned Domain Adaptation for Pedestrian Detection in Thermal Imagery", *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

et al., 2019, 2020a; Vandersteegen et al., 2018). A growing number of works have also investigated multispectral detectors that combine visible and thermal images for robust pedestrian detection (Wagner et al., 2016; Jingjing et al., 2016; Konig et al., 2017; Xu et al., 2017; Brazil et al., 2017a; Guan et al., 2018; Li et al., 2018, 2019).

However, multispectral detectors, in order to make the most out of both modalities, typically need to resort to additional (and expensive) annotations, and are usually based on far more complex network architectures than single-modality methods (see table 5.3). Moreover, due to the cost of deploying multiple aligned sensors (thermal and visible) at inference time, multispectral models can have limited applicability in real-world applications. Aside from the technical and economic reasons, the privacy-preserving affordances offered by thermal imagery are also a motivation for prefering thermal-only detecion (Kieu et al., 2019). Because of this, several recent works do not use visible images, but focus only on thermal images for pedestrian detection (John et al., 2015; Herrmann et al., 2018; Baek et al., 2017; Devaguptapu et al., 2019; Kieu et al., 2019; Guo et al., 2019; Kieu et al., 2020a). They typically yield lower performance than multispectral detectors since robust pedestrian detection using only thermal data is nontrivial and there is still potential for improvement.

There are a few task-conditioning approaches, such as conditional generative models like those based on adversarial networks (Mirza and Osindero, 2014) and the seminal work by Radford et al. (2015) that proposed architecture guidelines for training Deep Convolutional GANs. In particular, our approach is inspired by the general conditioning layer called Feature-wise Linear Modulation (FiLM) proposed by Perez et al. (2017) for conditioning visual reasoning tasks.

In this chapter we perform pedestrian detection on thermal imagery using a task-conditioned network architecture for domain adaptation. Our key idea is to augment a detector with an auxiliary network that solves a simpler classification task and then to exploit the learned representation of this auxiliary network to inject conditioning parameters into strategically chosen convolutional layers of the main detection network. Our method is based on the single-stage detector YOLOv3 (Redmon and Farhadi, 2018), whose computational efficiency makes it particularly well-suited to practical applications with real-time requirements. We extend the YOLOv3 architecture by integrating conditioning layers to better specialize the network to deal with day- and nighttime images. We evaluate conditioning of residual groups, detection heads, and their combination during domain adaptation. The resulting, adapted network operates entirely in the thermal domain and achieves excellent performance compared to other single-modality approaches.

The contributions of this work are:

- we propose a novel task-conditioned network architecture based on YOLOv3 (Redmon and Farhadi, 2018) that uses the auxiliary task of day/night classification to aid adaptation to the thermal domain;

- we conduct extensive ablative analyses probing the effectiveness of various task-conditioning architectures and adaptation schedules;
- to the best of our knowledge, our task-conditioned detection networks outperform all single-modality detection approaches on the KAIST Multispectral Pedestrian Detection Benchmark (Hwang et al., 2015); and
- exploiting only thermal imagery, we outperform many state-of-the-art multispectral pedestrian detectors on the KAIST benchmark at nighttime.

The rest of the chapter is organized as follows. In the next section we describe our approach to conditioning thermal domain adaptation on the auxiliary task of day/night discrimination. We report in section 5.3 on an extensive set of experiments performed to evaluate the effectiveness of task-conditioning, and in section 5.4 we compare with state-of-the-art results. In section 5.5 we conclude with a discussion of our contribution.

5.2 Task-conditioned Domain Adaptation

In this section we describe our approach to conditioning a detector during adaptation to the thermal domain. Our central idea is that robust pedestrian detection naturally depends on low-level semantic qualities of input images – for example whether an image is captured during the day or at night. This auxiliary information should be useful for learning representations upon which we can condition the adaptation internal representations used for the primary detection task. In the next section we describe the architecture of an auxiliary classification network that is connected to the main detection network, and in section 5.2.2 we describe the conditioning layers that can be strategically inserted into the network to modify internal representation. We describe two alternative conditioning architectures for YOLOv3 in section 5.2.3, and in section 5.2.4 we put everything together into a description of the combined adaptation loss.

5.2.1 Auxiliary classification network

Let $D_{\Theta_d}(\mathbf{x})$ represent the detector network (YOLOv3 in our case) parameterized by Θ_d, and let $F_i(\mathbf{x})$ represent the output of the i^{th} convolutional layer of the detection network. We define an auxiliary classification network as follows. The output of an early convolutional layer (e.g., $F_4(\mathbf{x})$ as in figure 5.1), is average pooled to form a feature that is then fed to two fully-connected layers of size C with ReLU activations. The resulting feature representation is then passed to a final fully connected layer with a single output and a sigmoid activation. We denote the output of this auxiliary network $A_{\Theta_a}(\mathbf{x})$.

During training we use the following loss attached to the output of the auxiliary network:

$$\mathcal{L}_a(\mathbf{x}_i, y_i; \Theta_a) = [y_i \cdot \log f(x_i) + (1 - y_i) \cdot \log(1 - f(x_i))], \tag{5.1}$$

where for all training images $\mathbf{x}_i$ we associate an auxiliary training label y_i. Since we experiment on the KAIST dataset, which distinguishes daytime and nighttime images in its annotations and evaluation protocol, we define $y_i = 0$ if $\mathbf{x}_i$ was captured during the day, and $y_i = 1$ if $\mathbf{x}_i$ was captured at night. In this case the auxiliary network has the task of classifying images as daytime or nighttime.

5.2.2 Conditioning layers

Our idea to use the internal, C-dimensional representation learned in the auxiliary classification network (i.e. the representation after the two fully-connected layers used for classification) rather than its output. See figure 5.1 for a schematic representation of the conditioning process. This representation is task-specific: in our experiments it is learned to capture the salient information *useful* for determining whether an image was captured during the day or at night. At strategic points in the main detection network we will use this representation to generate *conditioning parameters* that condition a convolutional feature map using the representation learned by the auxiliary network.

Consider an arbitrary convolutional output $F_i(\mathbf{x})$ of the main detector network D_{Θ_d}, and let d_i be the number of convolutional feature maps in $F_i(\mathbf{x})$. We generate conditioning parameters γ_i and β_i:

$$\begin{aligned} \gamma_i &= \text{ReLU}[W^i_\gamma A(\mathbf{x}) + b^i_\gamma] \\ \beta_i &= \text{ReLU}[W^i_\beta A(\mathbf{x}) + b^i_\beta], \end{aligned}$$

where $W^i_\gamma, W^i_\beta \in \mathbb{R}^{d_i \times C}$ and $b^i_\gamma, b^i_\beta \in \mathbb{R}^{d_i}$ are the weights and biases, respectively, of two new fully connected layers of D units added to the network (purple layers in Figure 5.1). These new layers are responsible for generating the parameters used for conditioning F_i.

F_i is substituted by the conditioned version:

$$F'_i(\mathbf{x}) = \text{ReLU}[(1 - \gamma_i) \odot F_i(\mathbf{x}) \oplus \beta_i],$$

where $\odot$ and $\oplus$ are, respectively, the elementwise multiplication and addition operations *broadcasted* to cover the spatial dimensions of the feature maps $F_i(\mathbf{x})$. In this way, the generated γ_i parameters can scale feature maps independently and the β_i parameters independently translate them.

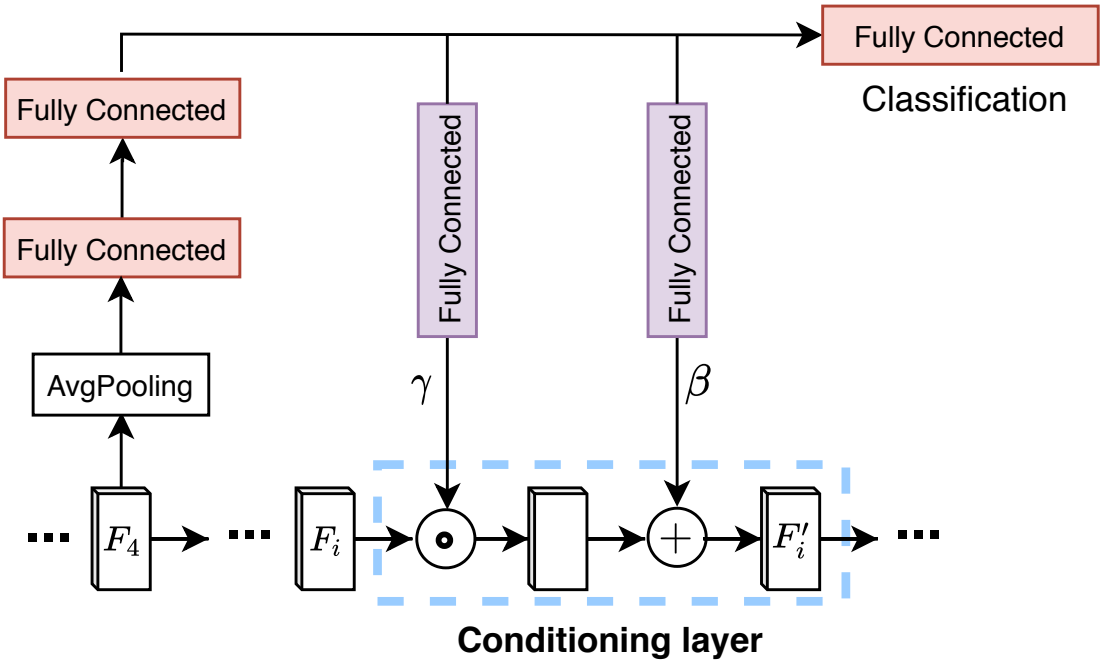


Figure 5.1: Conditioning layer and auxiliary classification network. The auxiliary network learns an internal representation used to solve a classification task. This representation is then leveraged by conditioning layers to adjust internal convolutional feature maps in the detection network.

5.2.3 Conditioned network architectures

YOLOv3 is a very deep detection network with three detection heads for detecting objects at different scales (Redmon and Farhadi, 2018). In order to investigate the effectiveness of conditioning YOLOv3 during domain adaptation, we experimented with two different strategies for injecting conditioning layers into the network. In section 5.3.3 we report on a series of ablation experiments performed to evaluate these different architectural possibilities for conditioning the network.

Conditioning residual groups (TC Res Group). YOLOv3 uses a 52-layer, fully-convolutional residual network as its backbone. The network is coarsely structured into five residual *groups*, each consisting of one or more residual blocks of two-convolutional layers with residual connections adding the input of each block to the output.

A natural conditioning point is at each of these residual groups. This strategy is illustrated in figure 5.2; the figure reports also the size of the layers of the conditioning network ($C = 1024$). After each group of residual blocks, we insert a conditioning layer after the last convolutional layer and *before* the final residual connection of the group.

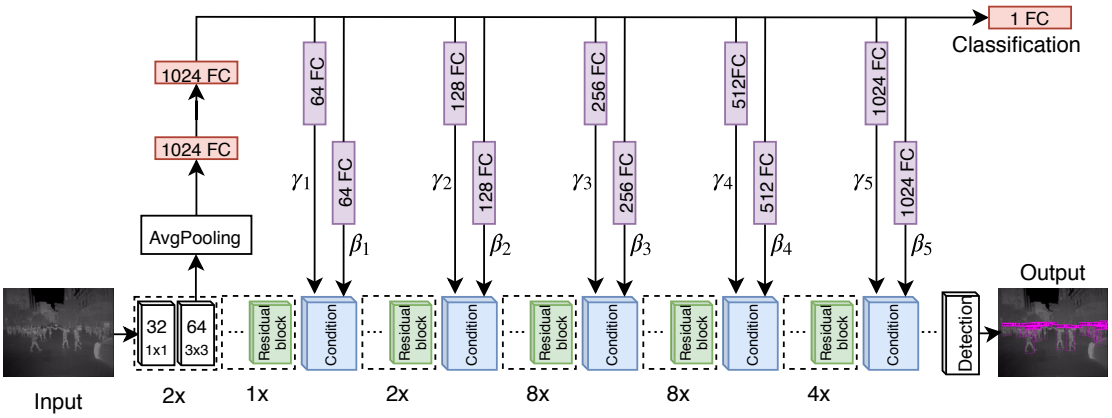


Figure 5.2: **TC Res Group**: Conditioning residual *groups* of YOLOv3. The pre-ReLU activations of the last layer of each convolutional group are modified by parameters γ_i and β_i. Conditioning is done before the final residual connection of each group.

Conditioning detection heads (TC Det). A natural alternative to conditioning residual groups is to condition each of the three detection heads branching off of the YOLOv3 backbone. The intuition here is to condition the network closer to where the actual detections are being made.

Detection heads in YOLOv3 consist of one convolutional block for the large-scale detection head, and three convolutional blocks for the other two. We insert the conditioning layer after the last convolution of these blocks and before the final 1×1 convolutional layer producing the detection head output. Figure 5.3 gives a schematic illustration of detection head conditioning architecture, and reports the size of the layers of the conditioning network ($C = 512$).

5.2.4 Adaptation loss

The final loss function used for domain adaptation is:

$$\mathcal{L}(\mathbf{x}_i, \mathbf{y}_i, y_i; \Theta_D, \Theta_A) = \mathcal{L}_d(\mathbf{x}_i, \mathbf{y}_i) + \mathcal{L}_a(\mathbf{x}_i, y_i),$$

where $\mathbf{x}$ is a training thermal image, $\mathcal{L}_d$ is the standard detection loss based on the structured target detections $\mathbf{y}_i$, and $\mathcal{L}_a$ is the auxiliary classification loss defined in equation (5.1).

When we backpropagate error from the auxiliary loss $\mathcal{L}_a$ we are improving the internal representation of the auxiliary network A_{Θ_a}, making it better for classifying day/night. When we backpropagate error from the detection loss, we simultaneously improve the generated conditioning parameters (γ_i, β_i) and the internal representation in the YOLOv3 backbone. Our intuition is that this adapts feature maps to be *conditionable* on based on the representation learned in the auxiliary classification network.

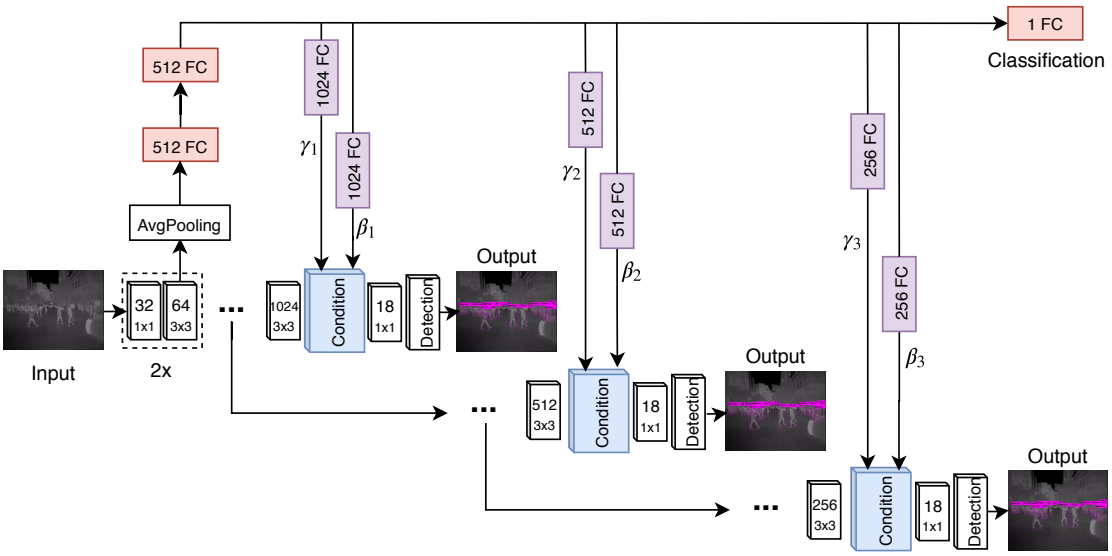


Figure 5.3: **TC Det**: Conditioning the detection heads of YOLOv3. Feature maps used for detection are conditioned using the internal representation of the auxiliary network.

5.3 Experimental Results

In this section we report results of a number of experiments we performed to evaluate the effectiveness of task-conditioned domain adaptation. In section 5.3.1 we describe the characteristics of the KAIST Multispectral Pedestrian Detection benchmark, and in section 5.3.3 we present two ablation studies we conducted to evaluate the various architectural parameters of our approach. In section 5.4 we compare with state-of-the-art single- and multimodal pedestrian detection approaches.

5.3.1 Dataset and evaluation metrics

Our experiments were conducted on the KAIST Multispectral Pedestrian Benchmark dataset (Hwang et al., 2015). KAIST is the only large-scale dataset with well-aligned visible/thermal pairs (Devaguptapu et al., 2019), and it contains videos captured both during the day and at night.

The KAIST dataset consists of 95,328 aligned visible/thermal image pairs split into 50,172 for training and 45,156 for testing. As is common practice, we use the *reasonable* setting (Dollar et al., 2012; Hwang et al., 2015; Kieu et al., 2019, 2020a), and use the improved training annotations from Li et al. (2018) and test annotations from Jingjing et al. (2016). We sample every two frames from training videos and exclude heavily occluded and small person instances (< 50 pixels). The final training set contains 7,601 images. The test set contains 2,252 image pairs sampled every 20 frames. Figure 5.4 shows some example images with our detection results

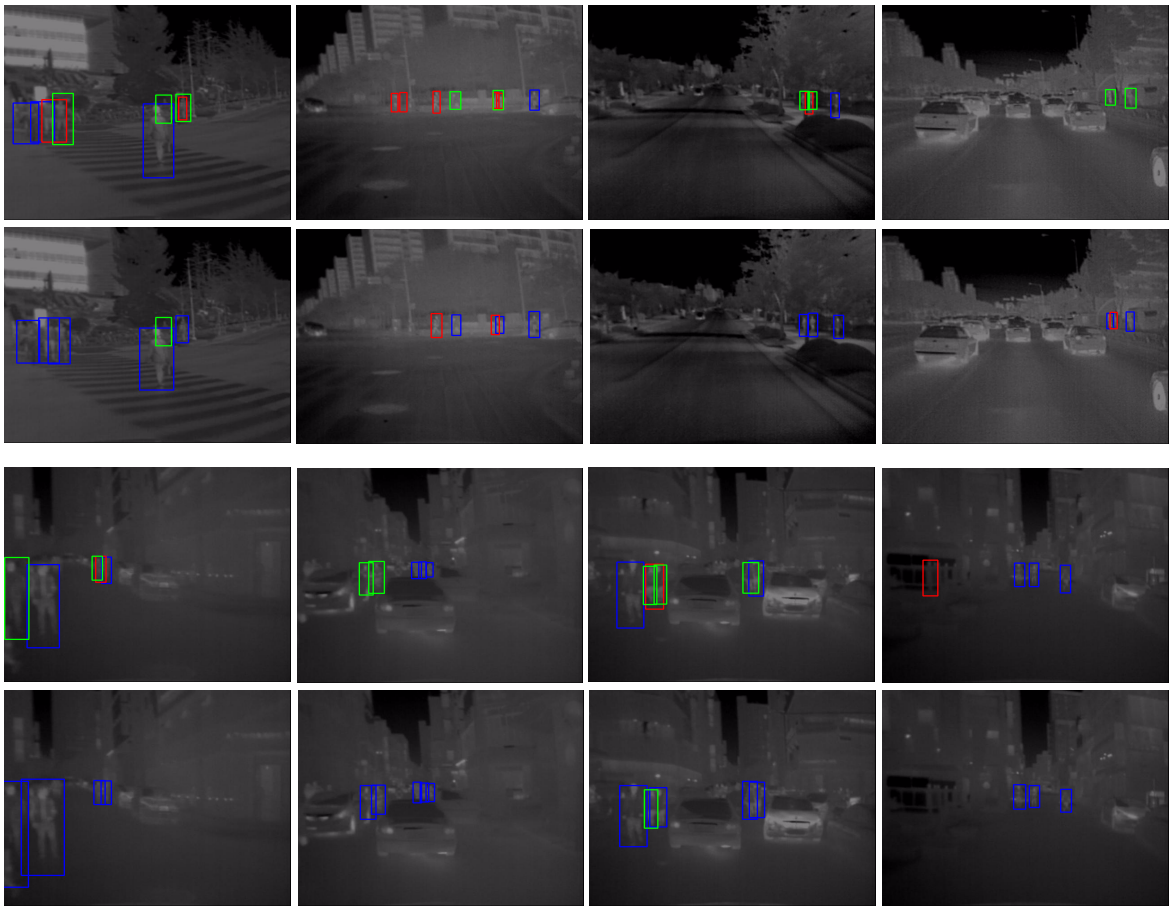


Figure 5.4: Examples of KAIST thermal images with detections. The first two columns are daytime images and the last two are nighttime. The first and the third rows show detection results without conditioning, and the second and last rows are detections with our **TC Det** detector. Blue boxes are true positive detections, green boxes are false negatives, and red boxes indicate false positives. See section 5.3.3 for detailed analysis.

on KAIST. The FLIR dataset was not used for this experiment because there are no state-of-the-art multispectral methods on FLIR dataset for comparison.

We used standard evaluation metrics for object detection, namely miss rate as a function of False Positives Per Image (FPPI), and log-average miss rate for thresholds in the range of $[10^{-2}, 10^{0}]$. For computing miss rates, an Intersection over Union (IoU) threshold of 0.5 is used to calculate True Positive (TP), False Positives (FP) and False Negatives (FN).

5.3.2 Implementation and training

All of our networks were implemented in PyTorch and source code and pretrained models are available.* During training, at each epoch we set aside 10% of the training images for validation for that epoch. We use the same hyperparameter settings of the original YOLOv3 model (Redmon and Farhadi, 2018) and use weights pretrained on MS COCO (Lin et al., 2014) as a starting point. We use Stochastic Gradient Descent (SGD) with an initial learning rate of 0.0001. When the validation performance no longer improves, we reduce the learning rate by a factor of 10. Training is halted after decreasing the learning rate twice in this way. All models were trained for a maximum of 50 epochs with a batch size of 8 and input image size 640×512. For most cases, training stops at around 30 epochs and requires about 12 hours on an NVIDIA GTX 1080.

5.3.3 Ablation studies

In this section we report on a series of experiments we conducted to explore the design space for task-conditioned adaptation of a pretrained YOLOv3 detector to the thermal domain. We first consider the *where*-aspect of task-conditioning (i.e. at which points in the YOLOv3 architecture task-conditioning is most effective), and then consider the *when*-aspect of task conditioning by exploring the many possibilities of conditioning adaptation phases.

Comparison of conditioning points. YOLOv3 is a very deep network which presents many options for intervening with conditioning layers. It has 23 residual blocks, each consisting of two convolutional layers and one residual connection. These 23 residual blocks are organized into five groups as illustrated in figure 5.2. Inspired by the (Perez et al., 2017), in which the authors demonstrate that conditioning residual blocks can be effective, we performed an architectural ablation on *where* to condition the network by considering conditioning of all residual blocks versus conditioning each residual group. We investigate also conditioning of the three detection heads, both alone and in combination with residual group conditioning.

The configurations investigated are:

- **No Conditioning** (direct fine-tuning on thermal): the YOLOv3 network pretrained on MSCOCO is directly fine-tuned on KAIST thermal images.

- **TC Res Group** (conditioning of residual groups): the conditioning scheme described in section 5.2.3 and illustrated in figure 5.2. We insert conditioning layers into all residual groups at the final residual block.

*`https://github.com/mrkieumy/task-conditioned`

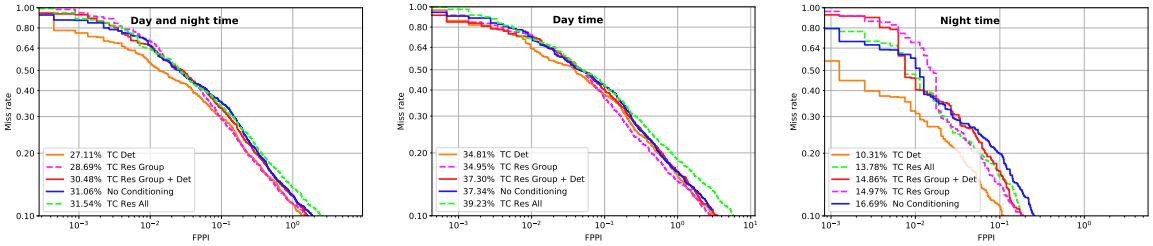


Figure 5.5: Ablation study of different conditioning points. Plots report miss rate as a function of false positives per image, and log-average miss rates are given in the legends.

- **TC Res All** (conditioning of all residual blocks): similar to group conditioning, but conditioning all residual blocks of the YOLOv3 network.
- **TC Det** (conditioning of detection heads): the scheme described in section 5.2.3 and illustrated in figure 5.3.
- **TC Res Group + Det** (conditioned residual groups and detection heads): a combination of **TC Res Group** and **TC Det**.

In figure 5.5 we plot the miss rate as a function of False Positive Per Image (FPPI) for the five different conditioning options. Note that most of the task-conditioned networks result in improvement over the **No Conditioning** network trained with standard fine-tuning. **TC Det** performs best overall and performs especially well at nighttime with a miss rate of only 10.31% – an improvement of 6.38% over the **No Conditioning** network.

While conditioning residual groups (**TC Res Group**) is also effective compared to fine-tuning, adding more conditioning layers results in worse performance. One reason for this might be that conditioning layers add parameters to the network, and depending on the size of the feature maps being conditioned could be leading to overfitting on the KAIST training set.

In figure 5.4 we give example detections from the **TC Det** and **No Conditioning** detectors. **TC Det** yields more true positive and fewer false positive detections with respect to simple fine-tuning. On daytime images (first two columns of figure 5.4), the detector without conditioning (top row) produces a number of false positives and missed detections which **TC Det** does not. The difference is even more pronounced at nighttime (second two columns of figure 5.4).

This ablation analysis indicates that conditioning *only* detection layers (**TC Det**) is most effective when compared to conditioning of residual blocks – answering the *where* of task-conditioning. In all of the following experiments we consider only the **TC Det** task-conditioned network.

Table 5.1: Ablation on adaptation schedules for **TC Det**. Results are on KAIST in terms of log-average miss rate (lower is better). **NC** indicates the modality is used for adaptation with no conditioning, **C** indicates the modality is used with conditioning of detection heads, and ✗ indicates the modality is not used during adaptation.

Training		**Testing**		**Miss Rate**		
visible	thermal	visible	thermal	all	day	night
NC	✗	✓	✗	36.67	32.83	45.00
C	✗	✓	✗	34.73	**29.53**	46.09
✗	NC	✗	✓	31.06	37.34	16.69
NC	NC	✗	✓	30.50	37.45	15.73
C	NC	✗	✓	28.48	35.86	12.97
✗	C	✗	✓	29.95	38.16	12.61
NC	C	✗	✓	28.53	36.59	11.03
C	C	✗	✓	**27.11**	34.81	**10.31**

Comparison of conditional adaptation schedules. In this set of experiments we compare the many options of conditioning when adapting a pretrained detector from the visible to the thermal domain to answer *when* to condition the network on the new domain. Starting from a pretrained detector, we can fine-tune (with or without conditioning) on KAIST RGB images, then fine-tune (again with or without conditioning) on KAIST thermal images. In table 5.1 we give results of an ablation study considering all these possibilities. Adapting first using RGB images, rather than going directly to thermal, is generally useful. In fact, the best adaptation schedule is to fine-tune a conditioning network on visible spectrum images, and then fine-tune that conditioned network on thermal imagery – answering the *when* of task-conditioning.

Visualizing the effects of conditioning. Figure 5.6 illustrates the effect conditioning has on the feature maps of YOLOv3. The heatmaps in this figure were generated by averaging the convolutional feature maps input to the medium-scale detection head of YOLOv3 and superimposing this on the original thermal image. The third column is the average feature map of a non-conditioned thermal detector (TD), and the fourth and fifth columns are, respectively, the average feature maps before and after conditioning.

From the heatmaps in figure 5.6 we note that pedestrians show more contrast with the background in the task-conditioned feature maps for both daytime and nighttime. Also, the thermal detector without conditioning misses several pedestrians and produces one false positive at nighttime, while TC Det correctly detects these and does not produce false positive detections. Task-conditioning also helps

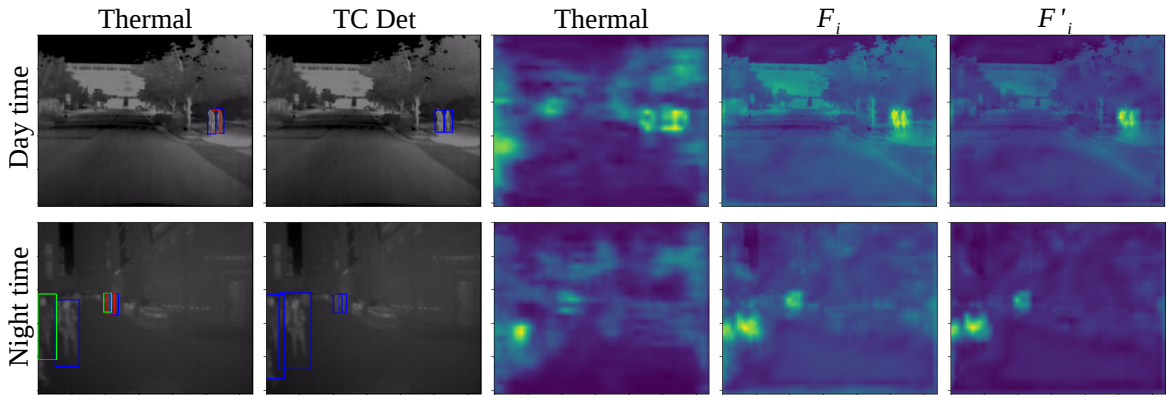


Figure 5.6: The effects of conditioning during daytime and nighttime. The first two columns show results for a thermal detector without conditioning and with conditioning. Blue boxes are true positive detections, green boxes are false negatives, and red boxes indicate false positives. See text detailed analysis.

eliminate one false positive in the daytime image.

Speed analysis. We also test the inference speed to compare between models. The average inference time for YOLOv3 is 28.57 milliseconds per image (∼35 FPS). Our **TC Det** network requires 33.17 milliseconds per image (∼30 FPS), and **TC Res Group** 35.01 milliseconds per image (∼29 FPS). Thus, task conditioning does not significantly increase the complexity of the network – in fact our **TC Det** network requires less than five milliseconds more for single-image inference compared to the original YOLOv3 detector.

5.4 Comparison with the State-of-the-art

In this section we compare our approaches with the state-of-the-art on KAIST. Since our approach focuses on detection only in thermal images at test time, we first compare with state-of-the-art single-modality detectors using only visible or only thermal images. Then, we compare our approaches with state-of-the-art multispectral detectors using both visible and thermal images.

5.4.1 Comparison with single-modality detectors.

Table 5.2 compares our approaches with the single-modality detectors using thermal-only or visible-only at training and testing time. **TC Visible** indicates the results of the second row in table 5.1, while **TC Thermal** and **TC Det** are the seventh and the last row in table 5.1, respectively. We can see that **TC Det** obtains the best results with a missrate of 27.11% in all modalities and 10.31% missrate at nighttime. Our

Table 5.2: Comparison with state-of-the-art single-modality approaches on KAIST in term of log-average miss rate (lower is better). Best results highlighted in **underlined bold**, second best in **bold**.

Detectors	MR all	MR day	MR night	Test images
FasterRCNN-C (Jingjing et al., 2016)	48.59	42.51	64.39	RGB
RRN+MDN (Xu et al., 2017)	49.55	47.30	54.78	RGB
FasterRCNN-T (Jingjing et al., 2016)	47.59	50.13	40.93	thermal
TPIHOG (Baek et al., 2017)	-	-	57.38	thermal
SSD300 (Herrmann et al., 2018)	69.81	-	-	thermal
Saliency Maps (Ghose et al., 2019)	-	**30.40**	21.00	thermal
VGG16-two-stage (Guo et al., 2019)	46.30	53.37	31.63	thermal
ResNet101-two-stage (Guo et al., 2019)	42.65	49.59	26.70	thermal
Bottom-up (Kieu et al., 2019)	35.20	40.00	20.50	thermal
Ours TC Visible	34.73	**29.53**	46.09	RGB
Ours TC Thermal	**28.53**	36.59	**11.03**	thermal
Ours TC Det	**27.11**	34.81	**10.31**	thermal

results outperform all existing single-modality methods by a large margin in all conditions (day, night, and all). To the best our knowledge, our detectors outperform all state-of-the-art single-modality approaches on KAIST dataset.

5.4.2 Comparison with multimodal detectors.

Table 5.3 compares our detectors with state-of-the-art multimodal approaches. **Ours Thermal** indicates for the result from the third row in table 5.1. Some multispectral methods using both visible and thermal images for training and testing such as MSDS (Li et al., 2018), IAF (Li et al., 2019) or IATDNN+IAMSS (Guan et al., 2018) are superior in terms of combined day/night miss rate (all). This is due to the advantage they have in exploiting both visible and thermal imagery, affecting in particular pedestrian detection during the day. In fact, the authors of MSDS (Li et al., 2018) proposed a set of manually "sanitized" annotations for KAIST that correct problems in the original annotations and their sanitized results at night-time (indicated by *) are better than the original results due to misalignment correction. Another key difference is that most state-of-the-art multispectral approaches use more complex, two-stage detection architectures like Faster RCNN (last column of table 5.3). Note, however, that TC Det result surpassed many multimodal techniques, and performs the best results at night.

We note that recent advances in the state-of-the-art on KAIST have been made by augmenting and/or correcting the original dataset annotations. For example, the authors of AR-CNN (Zhang et al., 2019b) completely re-annotated the KAIST dataset, correcting localization errors, adding relationships, and labeling unpaired

Table 5.3: Comparison with state-of-the-art multimodal approaches in terms of log-average miss rate on KAIST dataset (lower is better). All approaches use both visible and thermal at training and test time, while ours use only thermal imagery for testing. Results for Methods indicated with * were computed using detections provided by the authors. Best results highlighted in **underlined bold**, second best in **bold**.

Method		MR all	MR day	MR night	Detector Architecture
KAIST baseline	(Hwang et al., 2015)	64.76	64.17	63.99	ACF
Late Fusion	(Wagner et al., 2016)	43.80	46.15	37.00	RCNN
Halfway Fusion	(Jingjing et al., 2016)	36.99	36.84	35.49	Faster R-CNN
RPN+BDT	(Konig et al., 2017)	29.83	30.51	27.62	VGG-16 + BF
IATDNN+IAMSS	(Guan et al., 2018)	26.37	27.29	24.41	VGG-16 + RPN
IAF R-CNN*	Li et al. (2019)	20.95	21.85	18.96	Faster R-CNN
MSDS-RCNN	(Li et al., 2018)	**11.63**	**<u>10.60</u>**	13.73	VGG-16 + RPN
MSDS sanitized*	(Li et al., 2018)	**<u>10.89</u>**	**12.22**	**<u>7.82</u>**	VGG-16 + RPN
YOLO_TLV	(Vandersteegen et al., 2018)	31.20	35.10	22.70	YOLOv2
DSSD-HC	(Lee et al., 2018)	34.32	-	-	DSSD
GFD-SSD	(Zheng et al., 2019)	28.00	25.80	30.03	SSD
Ours Thermal		31.06	37.34	16.69	YOLOv3
Ours TC Res Group		28.69	34.95	14.97	YOLOv3
Ours TC Det		27.11	34.81	**10.31**	YOLOv3

objects, resulting in significantly improved performance. Use of additional manual annotations, however, renders their results impossible to compare with those of other approaches and are thus excluded from our comparison.

5.5 Conclusions

In this chapter we proposed a task-conditioned architecture for adapting visible-spectrum detectors to the thermal domain. Our approach exploits the internal learned representation of an auxiliary day/night classification network to inject conditioning parameters at strategic points in the detector network. Our experiments demonstrate that task-based conditioning of the YOLOv3 detection network can significantly improve thermal-only pedestrian detection performance.

Task-conditioned networks preserve the efficiency of the single-shot YOLOv3 architecture and perform respectably even compared to some multispectral detectors. However, they are outperformed by more complex, two-stage multispectral detectors such as MSDS (Li et al., 2018). We think, however, that our task-conditioning approach can also be fruitfully applied to such detectors by conditioning both region proposal and classification subnetworks.

Chapter 6

Generative synthesized thermal imagery for Domain Adaptation[†]

In this chapter, we propose a method for improving pedestrian detection in the thermal domain using two stages: first, a generative data augmentation approach is used, then a domain adaptation method using generated data adapts an RGB pedestrian detector. Our model, based on the Least-Squares Generative Adversarial Network, is trained to synthesize realistic thermal versions of input RGB images which are then used to augment the limited amount of labeled thermal pedestrian images available for training. We apply our generative data augmentation strategy in order to adapt a pre-trained YOLOv3 pedestrian detector to detect pedestrians in the thermal domain.

Experimental results demonstrate the effectiveness of our approach: using less than 50% of available real thermal training data, and relying on synthesized data generated by our model in the domain adaptation phase, our detector achieves state-of-the-art results on the KAIST Multispectral Pedestrian Detection Benchmark; even if more real thermal data is available adding GAN generated images to the training data results in improved performance, thus showing that these images act as an effective form of data augmentation. To the best of our knowledge, our detector achieves the best single-modality detection results on KAIST with respect to the state-of-the-art.

6.1 Introduction

Detectors based on thermal imagery have garnered attention recently as a means to mitigate the sensitivity of visible spectrum imagery to scene-incidental imaging con-

[†]Portions of this chapter were published in: M. Kieu, L. Berlincioni, L. Galteri, M. Bertini, A. D. Bagdanov, and A. Del Bimbo, "Robust pedestrian detection in thermal imagery using synthesized images." *Proceedings of the International Conference on Pattern Recognition (ICPR)*, 2020.

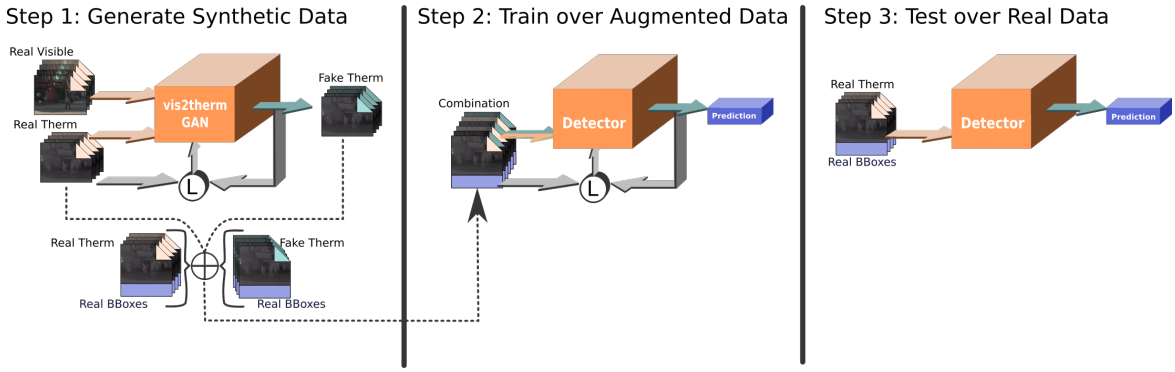


Figure 6.1: System overview: the vis2therm GAN generates fake thermal images from visible data; a mixture of real and fake thermal images along with related bounding boxes of objects are used to train an object detector, that is then tested on images from thermal cameras.

ditions (Kieu et al., 2019; Herrmann et al., 2018; Baek et al., 2017). However, thermal-only detectors typically yield lower performance than multispectral detectors since robust pedestrian detection using only thermal data is extremely challenging. A key performance-limiting factor is the relative lack of annotated thermal imagery available for training state-of-the-art models. Thermal pedestrian datasets are few, and – compared to visible-spectrum datasets – have orders of magnitude fewer annotated instances; for instance the Caltech Pedestrian Dataset (Dollar et al., 2012) has 350,000 annotations in the visible domain, while KAIST Multispectral Pedestrian dataset (Hwang et al., 2015) has $\sim$ 51,000 annotations and FLIR ADAS Dataset (FLIR, 2018) has $\sim$ 28,000. Scaling thermal-only detection to the levels of robustness and accuracy demanded by real-world applications is thus extremely difficult due to this poverty of annotated data.

Motivated by these challenges, in this chapter we propose to use a generative algorithm to perform data augmentation that can enrich thermal pedestrian datasets for training deep detector architectures. Our approach is based on a Least-Squares Generative Adversarial Network (LSGAN) (Mao et al., 2016) trained to synthesize thermal pedestrian images from RGB inputs. We investigate the best approaches to exploit these generated images during training, i.e. studying how to mix real thermal images with synthesized ones in order to effectively augment the training set. Experimental results indicate that our trained LSGAN is able to learn to translate RGB pedestrian images to useful thermal versions so that even using $\sim$ 50% synthetic images results in state-of-the-art pedestrian detection at nighttime and overall day/nighttime. This suggests that the approach can be extended to other domains in which thermal training data is scarce but is possible to effectively exploit the abundance of RGB imagery to adapt it to the thermal domain.

The contributions of this chapter are:

- we propose a novel generative model based on the Least-Squares Generative Adversarial Network (LSGAN) (Mao et al., 2016) that is able to synthesize thermal imagery from RGB;
- we propose a mixed real/synthetic training domain adaptation procedure that mixes real thermal imagery with thermal images synthesized from unlabeled RGB pedestrian images using our LSGAN and uses this augmented training set to adapt the YOLOv3 (Redmon and Farhadi, 2017) detector;
- we conduct extensive ablation study to probe the effectiveness of our approach and a variety of mixing proportions of real and synthesized imagery; and
- we conduct an extensive set of experiments comparing our approach to the state-of-the-art, and to the best of our knowledge our thermal-only detector outperforms all state-of-the-art single-modality detection approaches on the KAIST Multispectral Pedestrian Detection Benchmark (Hwang et al., 2015) by a large margin.

The rest of this chapter is organized as follows. In the next section, we review the scientific literature related to our proposed approach. In section 6.3 we describe our generative model used to synthesize thermal images and our training procedure used to adapt a YOLOv3 pedestrian detector to the thermal domain. We report in section 6.4 on an extensive set of experiments performed to evaluate the effectiveness of thermal pedestrian detection using our approach, and in section 6.5 we conclude with a discussion of our contribution.

6.2 Related Work

The problem of pedestrian detection in thermal imagery has attracted much attention from the research community over the years due to the advantages of thermal cameras in many real-world and critical applications.

6.2.1 Pedestrian detection in thermal imagery

Thanks to the reduction of costs and availability of multispectral cameras over the past few years, there are numerous recent works exploiting thermal images in combination with visible images for robust pedestrian detections as described in section 2 such as (Wagner et al., 2016) and (Jingjing et al., 2016) investigated many types of fusion of thermal and visible images; Konig et al. (2017) and Vandersteegen et al. (2018) composed RGB and thermal channel for multispectral pedestrian detection task; Xu et al. (2017) and Zhang et al. (2019a) learned the cross-modality

framework for multispectral task; Li et al. (2018) and Li et al. (2019) combined visible and thermal for two-branch network for investigating multispectral pedestrian detection task. In contrast, many recent works have investigated pedestrian detection using thermal (IR) imagery only. For example, John et al. (2015) used Adaptive fuzzy C-means for IR image segmentation and a CNN for pedestrian detection. Baek et al. (2017) proposed a combination of Thermal Position Intensity Histogram of Oriented Gradients (TPIHOG) and the additive kernel SVM (AKSVM) for nighttime-only detection in thermal imagery. Thermal images augmented with saliency maps, used as attention mechanism, have been used by Ghose et al. (2019).

The idea of performing several video preprocessing steps to make thermal images look more similar to grayscale images converted from RGB was investigated in (Herrmann et al., 2018), who then applied a pretrained and fine-tuned SSD detector. Recently, Cao et al. (2019) designed dual-pass fusion block (DFB) and channel-wise enhance module (CEM) to improve the one-stage detector RefineDet, and proposed their ThermalDet detector for pedestrian detection in thermal imagery. Another recent single-modality work was the Bottom-up Domain Adaptation approach proposed in (Kieu et al., 2019) for pedestrian detection in thermal imagery. We also focus on the thermal-only detection problem. However, our approach is distinct in that we concentrate on domain adaptation via data augmentation during training using synthetic thermal data which is generated by a generative model trained on unlabeled data.

6.2.2 Spectrum transfer between visible and thermal

The generation of RGB images from the thermal images has been approached as a grayscale colorization task in several previous works such as (Limmer and Lensch, 2016) where deep multiscale CNNs are used along with classical computer vision post processing techniques over near infrared images. In (Berg et al., 2018) a CNN is used with a more sophisticated objective function in order to tackle misalignment issues between the two visible and thermal modalities. In (Dong et al., 2018) instead an encoder-decoder architecture is applied for performing colorization.

Most recent works, however, rely heavily on generative models to perform image-to-image translation between visible and thermal. As defined in (Isola et al., 2017), the *image-to-image translation* problem is the task of translating one visual representation of a scene into another.

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014), are one the most significant recent improvements in the field of generative models and have been extensively used for image-to-image translation. The key feature of these models is the competitive min/max game between two networks. GANs have been successfully applied in many computer vision tasks such as super resolution (Galteri et al., 2017; Ledig et al., 2017; Wang et al., 2018b), style transferZhu et al.

(2017b), image in painting (Yeh et al., 2016) and domain adaptation (Hoffman et al., 2018).

Both Suarez et al. (2018) and Mehri and Sappa (2019) use GANs architectures to perform infrared and grayscale colorization. In (Suarez et al., 2018) a DCGAN with one seperate generator per channel is used, while in (Mehri and Sappa, 2019) an improved GAN (Zhu et al., 2017b) is proposed. Zhang et al. (2019) leverage multiple streams of polarimetric images to synthesize photo-realistic visible images of faces preserving discriminative features. In (Perera et al., 2017) a multi-image to image generative framework is presented, and one of the proposed settings is infrared and grayscale colorization. The use of these frameworks to perform data augmentation in order to improve the performance of a seperate classifier has been studied in multiple previous works such as (Antoniou et al., 2018) in which they focus on improving one-shot learning, in (Bowles et al., 2018) where segmentation of medical images is enhanced by GAN augmented data.

In this chapter we focus on the opposite task: mapping RGB images to the infrared spectrum. The closest related works are (Wang et al., 2019; Zhang et al., 2019; Guo et al., 2019; Kniaz et al., 2019), as they all employ generative models to translate images from the visible to the thermal spectrum. A modified Cycle-GAN (Zhu et al., 2017b) is used in (Wang et al., 2019), where the performance of drone detection in the thermal spectrum is improved using augmented data coming from a visible to thermal GAN framework, and also in (Guo et al., 2019), where a pedestrian detector is trained on augmented thermal data. Also in (Wang et al., 2019) a modified version is proposed which changing the loss with a perceptual texture loss term. In (Zhang et al., 2019), both pix2pix (Isola et al., 2017) and Cycle-GAN are used to generate thermal images to train an object tracker in the thermal domain; experiments show that images generated with pix2pix are of higher quality, since this approach operates on paired thermal/RGB data.

Kniaz et al. (2019) presented a framework for cross-modality color to thermal person re-identification. The generative model in this work is tasked with the generation of multiple thermal versions of the visible input image, which is then used to match with real thermal gallery set. Here the proposed architecture is a variation of (Zhu et al., 2017a), a multimodal image-to-image translation framework composed of multiple networks: cVAE-GAN from Larsen et al. (2016) and cLR-GAN from Chen et al. (2016) which are jointly optimized in a hybrid model in order to cover complementary tasks. One of the major contributions of Zhu et al. (2017a) is the ability to model the distribution of different correct outputs corresponding to the same input.

In our approach we instead rely on a different architecture that combines elements from (Mao et al., 2016) and (Wang et al., 2018b), as further detailed in Section 6.3.2. The ESRGAN architecture proposed by Wang et al. (2018b) focuses on

the *super-resolution* problem and improved over the previous state-of-the-art (Ledig et al., 2017) by introducing the Residual-in-Residual Dense Block, removing the Batch-Normalization layers, and changing the perceptual loss term.

6.3 Generative Data Augmentation for Thermal Domain Adaptation

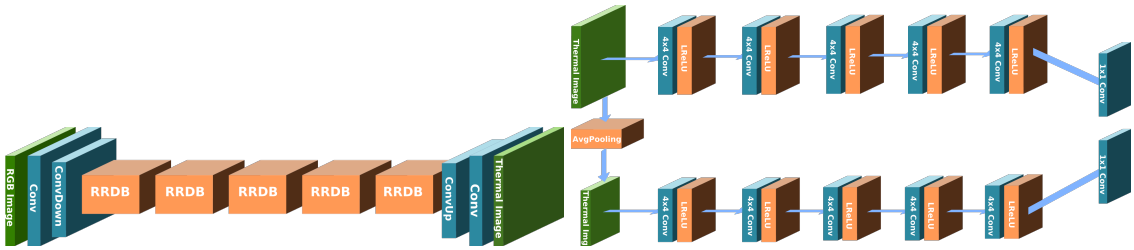
In this section we describe the two main components of our proposed approach. Our thermal pedestrian detector based on YOLOv3 (Redmon and Farhadi, 2018) is described in the next section, and our generative model which produces fake thermal images from available RGB images is described in section 6.3.2. An extensive series of experimental results are reported on in section 6.4.3.

6.3.1 Object detection in thermal images

We use YOLOv3 as our base pedestrian detector (Redmon and Farhadi, 2018). Following the Domain Adaptation approach described in (Kieu et al., 2019), we first adapt YOLOv3 in the visible domain by directly fine-tuning it on the visible spectrum images from the KAIST dataset (Hwang et al., 2015). Then, we use this detector as a starting point for training a thermal detector using a range of mixtures of real and GAN-generated thermal images. Figure 3.1 illustrates the original YOLOv3 architecture with thermal image as input and the output of the model at three detection scales.

We consider the following training regimes for thermal detectors:

- **Real-Thermal detector**: We directly fine-tune the detector on all available *real thermal images*.
- **Synthesized-Thermal detector**: We directly fine-tune the detector on all the *GAN-generated thermal images* (*synthesized images*).
- **Combined-Thermal detector**: We combine all available real images and all the synthesized images into a combined training set and then we fine-tune the detector on it. Note that the number of images in this combined set is double that used for the Real-Thermal and Synthesized-Thermal detectors.
- **Mixed-Thermal detectors**: We mix real images and synthesized images with a proportion varying from 10% to 90%; in total we have 9 mixed sets of images. For example, the mixed set 1 has 10% real images and 90% synthesized images. Note that the number of images used to train these detectors is the same as those used for Real-Thermal and Synthesized-Thermal detectors.



(a) Generator architecture (b) Discriminator architecture

Figure 6.2: Our LSGAN architecture. (a) is the proposed generator, and (b) the discriminator architecture composed of multiple Residual-in-Residual Dense Blocks and multiscale CNN.

For all experiments we evaluate performance on the KAIST test set of real thermal images.

6.3.2 Visible to thermal GAN

Our model is an LSGAN trained with both Adversarial and Perceptual losses. The Least Squares GAN (LSGAN) (Mao et al., 2016, 2017) improves on the standard GAN model by changing the loss function from a cross-entropy to a squared distance. It is comparatively more stable and easier to train. The Generator G architecture is built using the Residual in Residual Dense Block (RRDB) as the fundamental unit (see Figure 6.4). As in (Lim et al., 2017), we remove the batch normalization layer from the traditional *Conv-BN-LReLU* triplet. After the initial down-sampling convolutions five RRDB blocks are stacked in sequence as shown in Figure 6.2(a). Each RRDB block is composed of 4 Dense Blocks. Each Dense Block has a growth rate of $k = 32$ and contains five consecutive pairs of convolutional layers followed by a leaky rectified linear unit (LReLU) whose outputs are concatenated as shown in Figure 6.3.

Dense Blocks. DenseNets, introduced in (Huang et al., 2016), improve the information flow between layers by adding direct connections between a layer and all subsequent layers. By using this connectivity pattern the l^{th} layer receives the feature maps coming from all the preceding $l - 1$ layers as shown in Fig. 6.3. This dense connection strategy is realized by feeding as input the concatenation of every preceding layer output. DenseNets provide advantages both from a memory consumption and a vanishing gradient standpoint.

Residual in Residual Block. The composition of Residual Networks (He et al., 2016) and DenseNets is the Residual in Residual Dense block (RRDB), as introduced in (Wang et al., 2018b). A single RRDB is composed of multiple Dense blocks connected in a residual fashion, and is shown in Fig. 6.4. Finally, the output of the RRDB

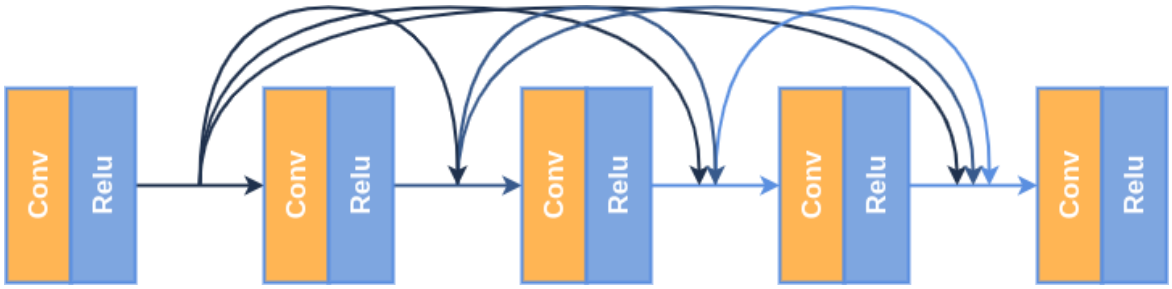


Figure 6.3: Dense Blocks. Arcs represent the concatenation between the output of a layer and one of its subsequent layers.

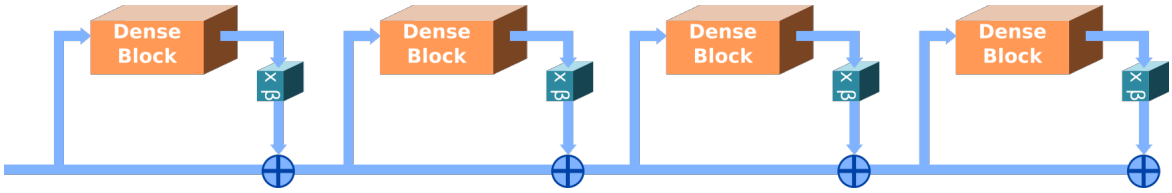


Figure 6.4: Residual in Residual Dense Block. The output of Dense Blocks are scaled by β and summed back to their input.

chain is followed by multiple *upscale-Conv-ReLU* blocks to scale the image back to input size.

Inspired by (Wang et al., 2018a; Durugkar et al., 2016; Karnewar and Wang, 2020) successful application of multi scale architectures we use a multi-scale discriminator D, shown in Figure 6.2(b), that makes no use of dense connectivity patterns. It is composed of five convolutional layers, each of them using a 4×4 convolutional kernel with stride 2 and followed by LReLU activation function. The number of feature maps is doubled as depth increases starting from 64. For each of the multiple scales, a single 1×1 convolutional filter is used as final output layer. Finally, the different outputs of every scale is evaluated independently.

Training. We trained the model as a Least Squares Generative Adversarial Network (LSGAN) with a perceptual loss. The discriminator D is trained as a standard LSGAN Discriminator:

$$\begin{aligned} L_{D_{LSGAN}} &= \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(D(x) - real_{label})^2] \\ &+ \frac{1}{2}\mathbb{E}_{z\sim p(z)}[(D(G(z)) - fake_{label})^2]. \end{aligned}$$

The generator loss is composed of three terms:

$$L_{G_{Adv}} = \frac{1}{2}\mathbb{E}_{z\sim p(z)}[(D(G(z)) - fake_{label})^2] \quad (6.1)$$

$$L_{G_{MAE}} = |real_{img} - fake_{img}| \quad (6.2)$$

$$L_{G_{Perceptual}} = (\phi^k(real_{img}) - \phi^k(fake_{img}))^2, \quad (6.3)$$

which are summed together:

$$L_{G_{LSGAN}} = L_{G_{Adv}} + L_{G_{MAE}} + L_{G_{Perceptual}} \quad (6.4)$$

Perceptual loss. Perceptual loss functions (Johnson et al., 2016) aim to provide a better measure for similarity compared to metrics such as the PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity Index). They have been shown useful for super-resolution and style-transfer tasks. Our perceptual loss architecture consists of two networks:

- Transformation Network T
- Loss Network ϕ

The Loss Network ϕ is pretrained as a classifier. When training the transformation network T, the loss network ϕ is used as a feature extractor and the distance between the target and the generated image in this feature space is used as a loss function for T. The main motivation behind perceptual loss functions lies in the intuition that computing distances in the high dimensional manifold extracted from a well-trained classifier should result in a better estimate compared to any pixel-space distance measure. As shown in Dosovitskiy and Brox (2016), pixel-space metrics can lead to minima that corresponds to blurry results. In this work, since our goal is to detect pedestrians, we use the pretrained YOLOv3 detector as a transformation network T to drive the generation of images. Equation (6.3) is a *perceptual loss* defined as the squared distance between the outputs ϕ^k of the k^{th} layer of a pretrained YOLOv3 network for a real and a generated input. We trained the ϕ network on KAIST for detection in a thermal images. We choose the last convolutional layer of YOLOv3 as representation of the input image in the high dimensional space learned by the classifier. Note that the loss network ϕ at this stage acts as a feature extractor and its weights are frozen.

6.4 Experimental Results

In this section we report on a range of experiments conducted to evaluate the effectiveness of our approach to thermal domain adaptation for pedestrian detection.

We first describe the dataset and evaluation metrics used, then in Section 6.4.2 give a qualitative evaluation of the performance of our GAN in generating thermal imagery from RGB input. In Section 6.4.3 we perform an ablative analysis of the use of synthetically generated thermal imagery for data augmentation, and in Section 6.4.4 give a comparison with the state-of-the-art.

6.4.1 Dataset, metrics, and experimental setup

Dataset. All of our experiments were conducted on the KAIST Multispectral Pedestrian Benchmark dataset (Hwang et al., 2015). KAIST is a large-scale dataset with well-aligned visible/thermal pairs (Devaguptapu et al., 2019), and it contains videos captured both during the day and at night. KAIST dataset consists of 95,328 image pairs split into 50,172 for training and 45,156 for testing. We follow the standard sampling procedure in (Hwang et al., 2015; Jingjing et al., 2016; Li et al., 2018), we sample every two frames from training videos and exclude heavily occluded and small person instances (< 50 pixels). The final training set contains 7,601 images. The test set contains 2,252 image pairs sampled every 20 frames. For training and testing, we use the improved training annotations from Li et al. (2018) and test annotations from Jingjing et al. (2016). We also do not use the FLIR dataset for the experiment because the visible-thermal image pairs are not well aligned.

Performance metrics. As is common practice to compare with the state-of-the-art, we used standard evaluation metrics for object detection, namely miss rate as a function of False Positives Per Image (FPPI), and log-average miss rate for thresholds in the range of $[10^{-2}, 10^{0}]$ with an Intersection over Union (IoU) threshold of 0.5 under the *reasonable* setting (Dollar et al., 2012; Hwang et al., 2015; Jingjing et al., 2016; Li et al., 2018; Kieu et al., 2020b). The *reasonable* setting is composed of *day-time*, *night-time*, and *all* (*both day and night time*) sets of images. Figure 6.5 shows some example images with our detection results on KAIST dataset.

Fine-tuning. All of our detectors were implemented using PyTorch. During fine-tuning to adapt to the thermal domain, at each epoch we set aside 10% of the training images for validation for that epoch. We trained every detector using Stochastic Gradient Descent with the same procedure and hyperparameters: image size 640×512, batch size of 4, We set an initial learning rate of 0.001 if the training set contains 50% or more real images, otherwise we use a learning rate of 0.0001. During fine-tuning, we reduce the learning rate by a factor of 10 every 3 epochs, and training is halted after 10 epochs.

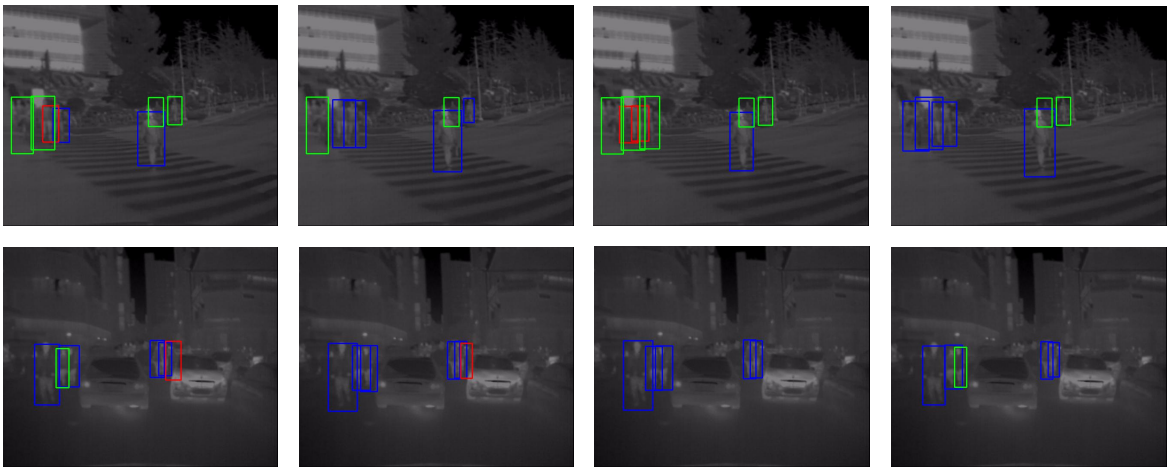


Figure 6.5: Examples of KAIST thermal images with detections. The first row is daytime images and the second is nighttime. The first and the second column are detection result on synthetic-only and real-only training, respectively. The third and the last column are combining all and mixed 90% proportion, respectively. Blue boxes are true positive detections, green boxes are false negatives, and red boxes indicate false positives. See section 6.4.4 for detailed analysis.

6.4.2 GAN results

The GAN framework for the *visible to thermal* transformation was trained on pairs of RGB-LWIR frames from the original training split of the KAIST dataset. In Figure 6.6 we show some examples detections using the detector trained with 20% sythesized images and 80% real images on two kinds of images. The first row shows detection results on generated images without Perceptual Loss $L_{G_{Perceptual}}$, and the second row gives detection results on generated images by our model trained with $L_{G_{Perceptual}}$. The use of the $L_{G_{Perceptual}}$ seems to result in more true positive (blue boxes) detection results, as well fewer false negative (green boxes).

6.4.3 Ablation study

In this section, we report on a series of experiments we conducted to explore the many options available when using GAN generated images (synthesized images) and thermal images (real images) for training the detectors described in Section 6.3.1. Initial experiments with simple augmentation strategies resulted in worse results than the conventional fine-tuning model. Rather than investigating how much synthetic images contribute to improving detection results, we focus on the potential of using fewer real thermal images with a small portion of synthesized thermal ones. This would be extremely useful for exploiting existing data on new domains where data is scarce.

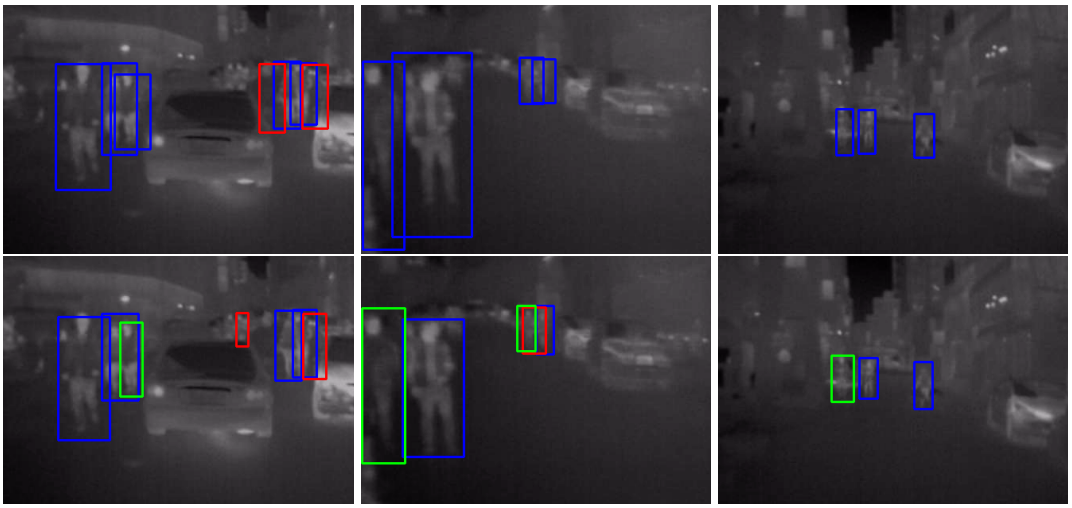


Figure 6.6: Example detections using the detector trained with 80% real images and 20% synthesized images. The first row shows detection results with the perceptual loss, while the second row is *without* perceptual loss. Blue boxes are true positive detections, green boxes are false negatives, and red boxes indicate false positives

Table 6.1: Ablation study on varying quantities of GAN-generated images. Results are on KAIST in terms of log-average miss rate (lower is better). Best results highlighted in **underlined bold**, second best in **bold**.

	Mixture		**Miss Rate (%)**		
	Real (%)	Synthetic (%)	all	day	night
Synthesized	0	100	45.88	54.37	26.04
Mixed	10	90	44.90	54.24	22.79
	20	80	41.21	51.04	18.92
	30	70	35.32	44.44	16.35
	40	60	34.78	43.45	14.53
	50	50	33.90	41.97	14.64
	60	40	31.50	39.83	12.33
	70	30	32.29	41.68	12.42
	80	20	**25.88**	**33.01**	**11.12**
	90	10	**25.62**	**31.86**	12.92
Real	100	0	28.46	36.32	**11.97**
Combined	all	all	34.29	41.93	16.80

Table 6.2: Comparison with state-of-the-art single-modality approaches on KAIST Thermal in term of log-average miss rate (lower is better). Best results highlighted in **underlined bold**, second best in **bold**.

Detectors		**MR all**	**MR day**	**MR night**
KAIST baseline	(Hwang et al., 2015)	64.76	64.17	63.99
FasterRCNN	(Jingjing et al., 2016)	47.59	50.13	40.93
TPIHOG	(Baek et al., 2017)	-	-	57.38
SSD300	(Herrmann et al., 2018)	69.81	-	-
Saliency + KAIST	(Ghose et al., 2019)	-	39.40	40.50
R^3-Net Saliency + KAIST	(Ghose et al., 2019)	-	**30.40**	21.00
VGG16-two-stage	(Guo et al., 2019)	46.30	53.37	31.63
ResNet101-two-stage	(Guo et al., 2019)	42.65	49.59	26.70
Bottom-up	(Kieu et al., 2019)	35.20	40.00	20.50
Ours Mixed 40_60		34.78	43.45	14.53
Ours Mixed 80_20		**25.88**	33.01	**11.12**
Ours Mixed 90_10		**25.62**	**31.86**	**12.92**

Thus, we use the conventional fine-tuning result as a baseline for comparison with various mixing strategies of GAN-generated thermal images. In table 6.1 we present results of an ablation study considering all these possibilities. From these results we first note that mixing in a *small* proportion of synthesized images (**Mixed**) rather than training on a all available real and synthesized images (**Combined**) is generally useful. In fact, the best mixture proportion is 90% real images with 10% percent synthesized images with 25.62% miss rate the “all" setting, and the second best is the **Mixed** of 80% and 20% with 11.12% miss rate in nighttime – an improvement of 5.68% over the **Combined** using all available data. Note that even with fewer than 50% real images our detector achieves results are comparable with state-of-the-art methods. Moreover, observe that mixing more than 50% real images results in improvement over the detector that combining all available real and synthesized images. The result reveals that the small portion of GAN synthesized images is useful for augmentation approach, but it must be consider based on the testing data such as the real test set was conducted on the test phase, thus the **Mixed** and **Real** results are better a little than the **Combined** result.

6.4.4 Comparison with the state-of-the-art

Table 6.2 compares our results with the state-of-the-art single modality approaches which are mostly trained and tested only on thermal images of KAIST dataset (except the KAIST baseline (Hwang et al., 2015) that is a multispectral method), some other models also used visible images for transfer learning such as (Kieu et al., 2019). We leveraged unlabeled RGB images of train set for generating synthetic thermal im-

ages, then we used this thermal data as augmentation for training; of course, testing was conducted on real thermal images of the test set. Results are compared in terms of log average miss rate (lower score is better). We can see that our approaches obtained the best results with 25.62% of missrate at “all" and 11.12% of missrate at “nighttime" – an improvement of 9.38% over the second state-of-the-art results. Moreover, our results outperform all existing the state-of-the-art methods by a large margin in both “night-time" and “all". The results of R^3-Net Saliency (Ghose et al., 2019) are a little better than ours in day time due to the advantages of their proposed pixel-level “saliency” annotation set with manually annotated 1,702 images from training and 369 from testing set, and their extraction of deep saliency maps by R^3-Net for augmenting thermal images of both training and testing.

Several different backbones have been used by the methods reported in the table, from VGG16 to Faster RCNN. Our backbone is the conventional YOLOv3 detector, and as fine tuning procedure we followed our previous approach of Kieu et al. (2019). The improvements that allowed to surpass the second-best state-of-the-art detector on KAIST (bottom-up (Kieu et al., 2019)) are: 1) the new data annotation as described in section 6.4.1; 2) the domain adaptation method of Kieu et al. (2019) and the experimentation with hyperparameter setting reported in section 6.4.1. Moreover, with the proposed generated synthesized thermal images with LSGAN and the mixed training procedure, we achieve state-of-the-art performance for both all (day and night) and nighttime.

It is expected that detection in thermal images at nighttime will always be better than daytime results because of the low contrast between pedestrians and background during the day, as noted in (Ghose et al., 2019).

In Figure 6.5 we show some example detections from four detectors (synthetics, real, combination and mixed90). From these examples we see that the mixed of 90% real images with 10% synthesized images yields more true positive and fewer false positive detections with respect to others. Not surprisingly, **synthesized detector** (the first column) produces a higher number of false positives and missed detections than **real detector** (the second column). The difference is even more pronounced at nighttime (second row of figure 6.5). The mixed scale 90% real with 10% synthesized images for training (the last columns) makes more true positive and less false positive than the **real detector**.

6.5 Conclusions

In this chapter we proposed a novel GAN architecture, based on LSGAN, to transform visible spectrum images in thermal spectrum ones. We also proposed a novel training procedure that mixes real and synthesized images to adapt the YOLOv3 detector for detection in the thermal domain. Extensive experimental validation

shows that our method outperforms state-of-the-art single-modality detectors for pedestrian detection on the KAIST dataset.

Our experiments show that that even using only 50% of available real thermal images it is possible to obtain results that are comparable with state-of-the-art methods trained using 100% real thermal images. This suggests that images generated with our proposed GAN are beneficial and may help to adapt visible spectrum detectors to operate in thermal spectrum in domains suffering from a lack of training data.

Chapter 7

Conclusions

In this dissertation we proposed four domain adaptation approaches for pedestrian detection in thermal imagery which our detectors outperform state-of-the-art, single-modality methods and are comparable with the best multispectral detectors on the KAIST Multispectral Pedestrian Benchmark.

To summarize the contributions of this work:

- **In Chapter 3** we proposed a bottom-up domain adaptation approach for pedestrian detection in thermal imagery and compared it with three top-down domain adaptation approaches based on fine-tuning. Our adaptation strategy is motivated by the fact that a thermal-only detectors better preserve privacy compared to visible spectrum or multispectral detectors. Our result promised the potential of bottom-up domain adaptation and the outcome of our work became the best paper award at the international conference on image analysis and processing.

- **In Chapter 4** by extending the advantages of our bottom-up adaptation method, we proposed a layer-wise domain adaptation approach which was validated on challenging thermal pedestrian detection datasets. The results reveal that a preliminary adaptation to visible spectrum images is useful to acquire domain knowledge that can be exploited after the final adaptation to the thermal domain. Exploiting only thermal domain, our results perform comparably with the state-of-the-art and outperform many multispectral approaches on KAIST. As far as we know, ours is the best performing detector on thermal imagery from the FLIR dataset.

- **In chapter 5** we proposed a task-conditioned network for domain adaptation which simultaneously solves two related tasks. The resulting detector is fast and robust in the thermal domain. Our task-conditioned detection network (TC-Det) achieves state-of-the-art results on the KAIST dataset. Our analysis

shows that task-conditioned networks exploit the internal learned representation of an auxiliary day/night classification sub-network to inject conditioning parameters at strategic points in the main detector network, and significantly improves pedestrian detection results on the KAIST dataset.

- **In chapter 6** generative data augmentation method for thermal domain adaptation which includes two stages: first, a Least-Squares Generative Adversarial Network is trained to synthesize realistic thermal versions of input RGB images which are then used to augment the limited amount of labeled thermal pedestrian images available for training. Then, we apply our generative data augmentation strategies in order to adapt a pre-trained RGB detector to detection in the thermal-only domain. By mixing a small portion of synthetic thermal data with real images, our detectors can achieve the best state-of-the-art single-modality results on the KAIST dataset. This suggests that images generated with our proposed GAN are beneficial and may help to adapt visible spectrum detectors to operate in thermal spectrum in domains suffering from a lack of training data.

Pedestrian detection is a problem that requires both high accuracy and real-time performance. There is still a trade-off between speed and accuracy. The best state-of-the-art, single-modality results with real-time speed (>24 FPS) still incur miss rate of about 10% – a number that must be improved for autonomous driving applications, for example. Multimodel methods can reach around 8% - 9% miss rate, however the speed is quite slow (<20 FPS). Closing the gap between multispectral pedestrian detectors and single-modality detectors in thermal imagery is a non-trivial task. There is still enormous potential to improve pedestrian detection results by using multispectral data to exploit only one domain. In particular, balancing the results between daytime and nighttime is crucial for single-modality models. Thermal imagery is inherently privacy-preserving, and we believe that thermal-only pedestrian detection has significant potential for the future if this balance can be found and the gap between single- and multi-modal detection closed.

Appendix A

Publications

Peer-reviewed international journals

1. **My Kieu**, Andrew D. Bagdanov, Marco Bertini, Alberto Del Bimbo, "Bottom-up and Layer-wise Domain Adaptation for Pedestrian Detection in Thermal Images", *ACM Transactions on Multimedia, Computing and Communications and Applications*, 2020.

2. **My Kieu**, Lorenzo Berlincioni, Marco Bertini, Andrew D. Bagdanov, Alberto Del Bimbo, "Thermal Image and Video Analysis and Understanding: A Review", *Journal of ACM Computing Surveys* (*ACM CSUR*), 2020 (in preparation).

Peer-reviewed international conferences

1. **My Kieu**, Andrew D. Bagdanov, Marco Bertini, Alberto Del Bimbo, "Domain Adaptation for Privacy-preserving Pedestrian Detection in Thermal Imagery", *International Conference on Image Analysis and Processing* (*ICIAP*), pages:203–213, 2019. (**Best Student Paper, Honorable Mention**).

2. **My Kieu**, Andrew D. Bagdanov, Marco Bertini, Alberto Del Bimbo, "Task-conditioned Domain Adaptation for Pedestrian Detection in Thermal Imagery", *European Conference on Computer Vision* (*ECCV*), 2020.

3. **My Kieu**, Lorenzo Berlincioni, Leonardo Galteri, Marco Bertini, Andrew D. Bagdanov, Alberto Del Bimbo, "Robust pedestrian detection in thermal imagery using synthesized images", *International Conference on Pattern Recognition* (*ICPR*), 2021.

Bibliography

Angelini, F., Yan, J., and Naqvi, S. (2019). Privacy-preserving online human behaviour anomaly detection based on body movements and objects positions. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*).

Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A., and Ferguson, D. (2015a). Real-time pedestrian detection with deep network cascades. In *Proc. of British Machine Vision Conference* (*BMVC*).

Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A., and Ferguson, D. (2015b). Real-time pedestrian detection with deep network cascades. In *BMVC*.

Antoniou, A., Storkey, A., and Edwards, H. (2018). Augmenting image classifiers using data augmentation generative adversarial networks. In *Proc. of Artificial Neural Networks and Machine Learning* (*ICANN*), pages 594–603.

Baek, J., Hong, S., Kim, J., and Kim, E. (2017). Efficient pedestrian detection at nighttime using a thermal camera. *Sensors*, 17(8):1850.

Benenson, R., Omran, M., Hosang, J., and Schiele, B. (2014). Ten years of pedestrian detection, what have we learned? In *Proc. of European Conference on Computer Vision* (*ECCV*).

Benenson Rodrigo, Omran Mohamed, H. J. and Bernt, S. (2014). Ten years of pedestrian detection, what have we learned? In *European Conference on Computer Vision - ECCV 2014 Workshops*.

Berg, A., Ahlberg, J., and Felsberg, M. (2018). Generating visible spectrum images from thermal infrared. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), pages 1224–122409.

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R. N., Hammers, A., Dickie, D. A., del C. Valdés Hernández, M., Wardlaw, J. M., and Rueckert, D. (2018). Gan augmentation: Augmenting training data using generative adversarial networks. *ArXiv*, abs/1810.10863.

Brazil, G., Yin, X., and Liu, X. (2017a). Illuminating pedestrians via simultaneous detection & segmentation. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*.

Brazil, G., Yin, X., and Liu, X. (2017b). Illuminating pedestrians via simultaneous detection and segmentation. *ICCV*, pages 4960–4969.

Cao, J., Pang, Y., Xie, J., Khan, F., and Shao, L. (2020). From handcrafted to deep features for pedestrian detection: A survey. *ArXiv*, abs/2010.00456.

Cao, Y., Guan, D., Wu, Y., Yang, J., Cao, Y., and Yang, M. Y. (2019). Box-level segmentation supervised deep neural networks for accurate and real-time multispectral pedestrian detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150:70–79.

Cao, Y., Zhou, T., Zhu, X., and Su, Y. (2019). Every feature counts: An improved one-stage detector in thermal imagery. In *Proc. of IEEE International Conference on Computer and Communications (ICCC)*, pages 1965–1969.

Chen, C., Dou, Q., Chen, H., Qin, J., and Heng, P.-A. (2019). Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of The Thirty-Third Conference on Artificial Intelligence (AAAI)*, pages 865–872.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 2180–2188.

Chen, Y., Jhong, S., Li, G., and Chen, P. (2019). Thermal-based pedestrian detection using faster r-cnn and region decomposition branch. In *2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 1–2.

Chen, Y.-Y., Li, G.-Y., Sin-Ye, J., Chen, P.-H., Tsai, C.-C., and Chen, P.-H. (2020). Nighttime pedestrian detection based on thermal imaging and convolutional neural networks. *Sensors and Materials*, 32:3157.

Chengyang Li, Dan Song, R. T. and Tang, M. (2018). Multispectral pedestrian detection via simultaneous detection and segmentation. *CoRR*.

Choi, Y., Choi, M.-J., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2017). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

Davis, J. and Sharma, V. (2007). Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106:162–182.

Davis, J. W. and Keck, M. A. (2005). A two-stage template approach to person detection in thermal imagery. In *2005 Seventh IEEE Workshops on Applications of Computer Vision* (*WACV/MOTION'05*) - *Volume 1*, volume 1, pages 364–369.

Devaguptapu, C., Akolekar, N., M Sharma, M., and N Balasubramanian, V. (2019). Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops* (*CVPR-W*).

Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2009). Pedestrian detection: A benchmark. *2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (*TPAMI*), 34(4):743–761.

Dong, Z., Kamata, S.-i., and Breckon, T. P. (2018). Infrared image colorization using a s-shape network. In *2018 25th IEEE International Conference on Image Processing* (*ICIP*), pages 2242–2246. IEEE.

Dosovitskiy, A. and Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. In *Proc. of Advances in Neural Information Processing Systems* (*NIPS*).

Dou, Q., Ouyang, C., Chen, C., Chen, H., and Heng, P.-A. (2018). Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 691–697. International Joint Conferences on Artificial Intelligence Organization.

Du, X., El-Khamy, M., Lee, J., and Davis, L. (2017). Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection. In *Proc. of IEEE Winter Conference on Applications of Computer Vision* (*WACV*).

Durugkar, I. P., Gemp, I., and Mahadevan, S. (2016). Generative multi-adversarial networks. *CoRR*, abs/1611.01673.

FLIR (2018). Flir starter thermal dataset.

Fritz, K., König, D., Klauck, U., and Teutsch, M. (2019). Generalization ability of region proposal networks for multispectral person detection. In *Proc. of Automatic Target Recognition XXIX*, volume 10988.

Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., and Berg, A. C. (2017). Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*.

Gade, R. and Moeslund, T. B. (2013). Thermal cameras and applications: a survey. *Machine Vision and Applications*, 25:245–262.

Galteri, L., Seidenari, L., Bertini, M., and Del Bimbo, A. (2017). Deep generative adversarial compression artifact removal. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*.

Gawande, U., Hajari, K., and Golhar, Y. (2020). Pedestrian detection and tracking in video surveillance system: Issues, comprehensive review, and challenges.

Ghiass, R., Arandjelovic, O., Bendada, H., and Maldague, X. (2014). Infrared face recognition: A comprehensive review of methodologies and databases. *Pattern Recognition*.

Ghose, D., Desai, S. M., Bhattacharya, S., Chakraborty, D., Fiterau, M., and Rahman, T. (2019). Pedestrian detection in thermal images using saliency maps. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Gonzalez Alzate, A., Fang, Z., Socarras, Y., Serrat, J., Vázquez, D., Xu, J., and López, A. (2016). Pedestrian detection at day/night time with visible and fir cameras: A comparison. *Sensors*, 16:820.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*.

Guan, D., Cao, Y., Yang, J., Cao, Y., and Yang, M. Y. (2018). Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157.

Guo, T., Huynh, C. P., and Solh, M. (2019). Domain-adaptive pedestrian detection in thermal images. In *Proc. of IEEE International Conference on Image Processing (ICIP)*.

Hadi, H., Mamat, R., Sheikh, U. U., and Amin, S. (2015). Fusion of thermal and depth images for occlusion handling for human detection from mobile robot. pages 1–5.

Hazan, A., Shoshan, Y., Khapun, D., Aladjem, R., and Ratner, V. (2018). Adapternet - learning input transformation for domain adaptation. *CoRR*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Heo, D., Lee, E., and Ko, B. (2017). Pedestrian detection at night using deep neural networks and saliency maps. *Journal of Imaging Science and Technology*, 61.

Herrmann, C., Ruf, M., and Beyerer, J. (2018). CNN-based thermal infrared person detection by domain adaptation. In *Proc. of Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*, volume 10643. International Society for Optics and Photonics.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., and Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. of International Conference on Machine Learning (ICML)*.

Huang, G., Liu, Z., and Weinberger, K. Q. (2016). Densely connected convolutional networks. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hwang, S., Park, J., Kim, N., Choi, Y., and Kweon, I. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jingjing, L., Shaoting, Z., Shu, W., and Dimitris, M. (2016). Multispectral deep neural networks for pedestrian detection. In Richard C. Wilson, E. R. H. and Smith, W. A. P., editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 73.1–73.13. BMVA Press.

John, V., Mita, S., Liu, Z., and Qi, B. (2015). Pedestrian detection in thermal images using adaptive fuzzy c-means clustering and convolutional neural networks. In *Proc. of IAPR International Conference on Machine Vision Applications (MVA)*, pages 246–249.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *Proc. of European Conference on Computer Vision (ECCV)*.

Karnewar, A. and Wang, O. (2020). Msg-gan: Multi-scale gradients for generative adversarial networks. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kieu, M., Bagdanov, A. D., and Bertini, M. (2020a). Bottom-up and layer-wise domain adaptation for pedestrian detection in thermal images. In *ACM Transactions on Multimedia Computing Communications and Applications*. Association for Computing Machinery (ACM).

Kieu, M., Bagdanov, A. D., Bertini, M., and Del Bimbo, A. (2019). Domain adaptation for privacy-preserving pedestrian detection in thermal imagery. In *Proc. of International Conference on Image Analysis and Processing (ICIAP)*.

Kieu, M., Bagdanov, A. D., Bertini, M., and Del Bimbo, A. (2020b). Task-conditioned domain adaptation for pedestrian detection in thermal imagery. In *European Conference on Computer Vision (ECCV)*. Springer.

Kniaz, V. V., Knyaz, V. A., Hladůvka, J., Kropatsch, W. G., and Mizginov, V. (2019). Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proc. of European Conference on Computer Vision Workshops (ECCV-W)*, pages 606–624, Cham. Springer International Publishing.

Konig, D., Adam, M., Jarvers, C., Layher, G., Neumann, H., and Teutsch, M. (2017). Fully convolutional region proposal networks for multispectral person detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Kouw, W. M. (2018). An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*.

Kristoffersen, M. S., Dueholm, J. V., Gade, R., and Moeslund, T. (2016). Pedestrian counting with occlusion handling using stereo thermal cameras. *Sensors (Basel, Switzerland)*, 16.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In *Proc. of International Conference on Machine Learning (ICML)*.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-

resolution using a generative adversarial network. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

Lee, Y., Bui, T. D., and Shin, J. (2018). Pedestrian detection based on deep fusion network using feature correlation. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference* (*APSIPA ASC*), pages 694–699.

Li, C., Song, D., Tong, R., and Tang, M. (2018). Multispectral pedestrian detection via simultaneous detection and segmentation. In *Proc. of British Machine Vision Conference* (*BMVC*).

Li, C., Song, D., Tong, R., and Tang, M. (2019). Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171.

Li, J., Liang, X., Shen, S., Xu, T., Feng, J., and Yan, S. (2017). Scale-aware fast R-CNN for pedestrian detection. *IEEE Transactions on Multimedia* (*TMM*), 20(4):985–996.

Li, J., Liang, X., Shen, S., Xu, T., and Yan, S. (2015). Scale-aware fast R-CNN for pedestrian detection. *CoRR*.

Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. (2017). Enhanced deep residual networks for single image super-resolution. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*).

Limmer, M. and Lensch, H. P. A. (2016). Infrared colorization using deep convolutional neural networks. In *Proc. of IEEE International Conference on Machine Learning and Applications* (*ICMLA*).

Lin, T., Maire, M., Serge J. Belongie, L. D. B., Girshick, R. B., James Hays, P. P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*.

Liu, L., Bao, H., Pan, W., and Xu, C. (2016a). Night-time pedestrian detection based on temperature and hogi feature in infrared images. *International Journal of Simulation–Systems, Science & Technology*, 17(22).

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., and Pietikäinen, M. (2019a). Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128:261–318.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016b). SSD: Single shot multibox detector. In *Proc. of European Conference on Computer Vision* (*ECCV*), pages 21–37. Springer.

Liu, W., Liao, S., Ren, W., Hu, W., and Yu, Y. (2019b). High-level semantic feature detection: A new perspective for pedestrian detection. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. (2015). Learning transferable features with deep adaptation networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 97–105. JMLR.org.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., and Wang, Z. (2016). Multi-class generative adversarial networks with the L2 loss function. *CoRR*, abs/1611.04076.

Markit, I. (2019). 245 million video surveillance cameras installed globally in 2014. Web page. Accessed: May 5, 2019.

Masana, M., van de Weijer, J., Herranz, L., Bagdanov, A. D., and Alvarez, J. M. (2017). Domain-adaptive deep network compression. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*.

Mehri, A. and Sappa, A. D. (2019). Colorizing near infrared images through a cyclic adversarial approach of unpaired samples. In *Proc. of CVPR-W*.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Nakashima, S., Kitazono, Y., Zhang, L., and Serikawa, S. (2010). Development of privacy-preserving sensor for person detection. *Procedia - Social and Behavioral Sciences*, 2:213–217.

Negied, N. K., Hemayed, E., and Fayek, M. B. (2015). Pedestrians' detection in thermal bands – critical survey. *Journal of Electrical Systems and Information Technology*, 2:141–148.

Olmeda, D., Premebida, C., Nunes, U., Armingol, J., and de la Escalera, A. (2013). Pedestrian detection in far infrared images. *Integrated Computer-Aided Engineering*, 20.

Ouyang, W., Zeng, X., and Wang, X. (2016a). Learning mutual visibility relationship for pedestrian detection with a deep model. *International Journal of Computer Vision (IJCV)*, 120(1):14–27.

Ouyang, W., Zeng, X., and Wang, X. (2016b). Learning mutual visibility relationship for pedestrian detection with a deep model. *International Journal of Computer Vision*, pages 14–27.

Perera, P., Abavisani, M., and Patel, V. (2017). In2i : Unsupervised multi-image-to-image translation using generative adversarial networks. In *Proc. of International Conference on Pattern Recognition (ICPR)*.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. C. (2017). FiLM: Visual reasoning with a general conditioning layer. In *Proc. of AAAI Conference on Artificial Intelligence (AAAI)*.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434.

Redmon, J. and Farhadi, A. (2017). Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *CoRR*.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*, pages 91–99.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Socarras, Y., Ramos, S., Vazquez, D., Lopez, A., and Gevers, T. (2013). Adapting pedestrian detection from synthetic to far infrared images. In *ICCV – Workshop on Visual Domain Adaptation and Dataset Bias*, Sydney, Australia.

Suarez, P., Sappa, A., and Vintimilla, B. (2018). Learning to colorize infrared images. In *Proc. of PAAMS*.

Teng, Y., Choromanska, A., and Bojarski, M. (2018). Invertible autoencoder for domain adaptation. *CoRR*.

Tian, Y., Luo, P., Wang, X., and Tang, X. (2015). Pedestrian detection aided by deep learning semantic tasks. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vandersteegen, M., Van Beeck, K., and Goedemé, T. (2018). Real-time multispectral pedestrian detection with a single-pass deep neural network. In *Proc. of International Conference Image Analysis and Recognition* (*ICIAR*).

Wagner, J., Fischer, V., Herman, M., and Behnke, S. (2016). Multispectral pedestrian detection using deep fusion convolutional neural networks. In *Proc. of 24th European Symposium on Artificial Neural Networks* (*ESANN*), pages 509–514.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018a). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Change Loy, C. (2018b). Esrgan: Enhanced super-resolution generative adversarial networks. In *Proc. of European Conference on Computer Vision* (*ECCV*).

Wang, Y., Chen, Y., Choi, J., and Kuo, C. J. (2019). Towards visible and thermal drone monitoring with convolutional neural networks. *APSIPA Transactions on Signal and Information Processing*, 8.

Wang, Z., Chen, Z., and Wu, F. (2018c). Thermal to visible facial image translation using generative adversarial network. *IEEE Signal Processing Letters*, PP:1–1.

Xu, D., Ouyang, W., Ricci, E., Wang, X., and Sebe, N. (2017). Learning cross-modal deep representations for robust pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pages 5363–5371.

Yeh, R. A., Chen, C., Lim, T., Hasegawa-Johnson, M., and Do, M. N. (2016). Semantic image inpainting with perceptual and contextual losses. *CoRR*, abs/1607.07539.

Yonglong Tian, Ping Luo, X. W. and Tang, X. (2014). Pedestrian detection aided by deep learning semantic tasks. *CoRR*.

Zhang, H., Riggan, B., Hu, S., Short, N., and Patel, V. (2019). Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *International Journal of Computer Vision* (*IJCV*), 127:1–18.

Zhang, L., Gonzalez-Garcia, A., van de Weijer, J., Danelljan, M., and Khan, F. S. (2019). Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing*, 28(4):1837–1850.

Zhang, L., Lin, L., Liang, X., and He, K. (2016). Is faster R-CNN doing well for pedestrian detection? In *Proc. of European Conference on Computer Vision* (*ECCV*).

Zhang, L., Liu, Z., Chen, X., and Yang, X. (2019a). The cross-modality disparity problem in multispectral pedestrian detection. *arXiv preprint arXiv:1901.02645*.

Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., and Liu, Z. (2019b). Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5127–5137.

Zheng, Y., Izzat, I. H., and Ziaee, S. (2019). GFD-SSD: Gated fusion double SSD for multispectral pedestrian detection. *arXiv preprint arXiv:1903.06999*.

Zhu, J., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., and Shechtman, E. (2017a). Toward multimodal image-to-image translation. In *Proc. of Advances in Neural Information Processing Systems (NIPS)*.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017b). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*.

Zou, Z., Shi, Z., Guo, Y., and Ye, J. (2019). Object detection in 20 years: A survey. *CoRR*, abs/1905.05055.